ugr | Universidad de Granada

TÉCNICAS ESTADÍSTICAS APLICADAS AL ESTUDIO DE DATOS EXPERIMENTALES

Aplicación de la técnica de Regresión Lineal Simple a la relación *Contribution – Quality* en el análisis de correspondencias en *data mining* con R.TeMiS [*R Text Mining Solution*]

Un estudio de caso: Regresión Lineal Simple aplicada al resultado del análisis factorial de correspondencias de los temas de investigación en *Translational research* y *Personalized medicine* según SCOPUS.

por

Jesús Minguillón-Campos [*] *y* ***José Pino-Díaz*** [**]

(*) Universidad de Granada; Departamento de Arquitectura y Tecnología de Computadoras; Centro de Investigación en Tecnologías de la Información y la Comunicación; C/ Periodista Rafael Gómez Montero, 2; 18014-Granada .

(**) Universidad de Málaga; Grupo de Investigación Techné Ingeniería del Conocimiento y del Producto; Campus de Fuentenueva, 18071-Granada.

Introducción

Investigación traslacional y medicina personalizada son líneas de investigación recientes. El concepto de medicina traslacional se identifica con el objetivo de facilitar la transición de la investigación básica en aplicaciones clínicas que redunden en beneficio de la salud (Wehling, M., 2008).

Investigación traslacional (*translational research*) no se debe confundir con investigación aplicada (*applied research*), por ejemplo en la industria farmacéutica la expresión *translational research* se refiere al traslado de los conocimientos de la investigación básica a la búsqueda de fármacos que curen las enfermedades. Estos estudios, de carácter preliminar, preceden a los ensayos clínicos a gran escala, propios de la investigación aplicada, etapa final de la investigación para el registro y comercialización de un medicamento. La investigación traslacional es una investigación básica aplicada a las primeras fases del desarrollo de un medicamento (Cabo Salvador, J.).

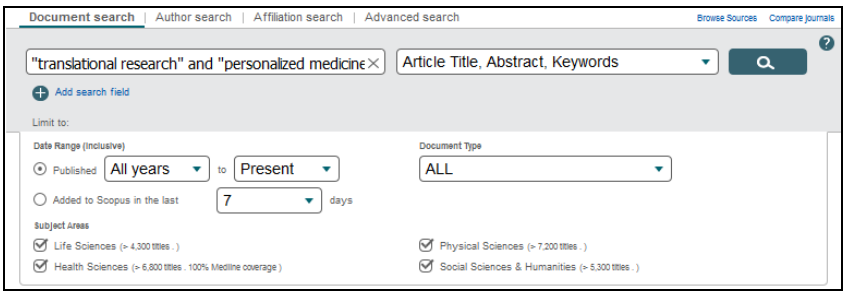
La medicina personalizada (*personalized medicine*) tiene como objetivo el tratamiento de cada individuo de manera específica atendiendo a diversos parámetros, como los genéticos o proteicos, que ayudan a mejorar los aspectos preventivos, diagnósticos y terapéuticos (Gamero Estévez, E.).

Así pues la investigación traslacional y la medicina personalizada son conceptos que se unen con el objetivo de acercar la investigación básica al paciente, aplicando los descubrimientos generados durante la investigación en el laboratorio y en estudios preclínicos al desarrollo de ensayos clínicos personalizados (Universidad de Granada).

Objeto

El presente trabajo tiene por finalidad aplicar la técnica de Regresión Lineal Simple (RLS) a los resultados del análisis factorial de correspondencias realizado a los términos y *clusters* resultantes de la clasificación de los temas de investigación en Investigación traslacional y Medicina personalizada. Se pretende estudiar la dependencia lineal de dos de los parámetros obtenidos en el análisis textual, *Contribution* y *Quality*. Para el análisis de correspondencias se han empleado técnicas de análisis de datos textuales y de estadística lexical con un corpus documental extraído de SCOPUS, base de datos internacional y multidisplinar.

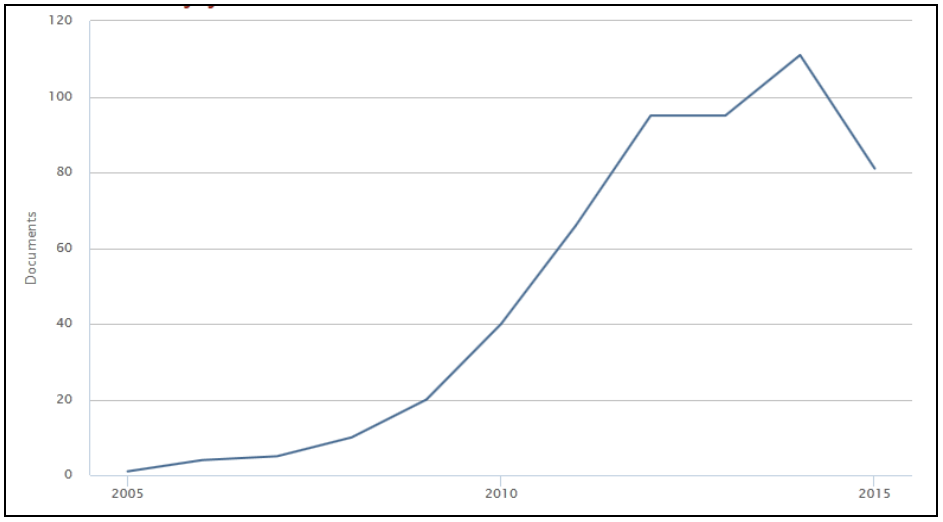
Los documentos del corpus proceden de una búsqueda sobre Investigación traslacional y Medicina personalizada empleando la ecuación de búsqueda ["*translational research*" *and* "*personalized medicine*"], realizada el 16 de diciembre de 2015.



Como resultado se han obtenido 526 documentos de todo tipo (artículos, revisiones, libros, etc).

Year	Documents
2015	81
2014	111
2013	95
2012	95
2011	66
2010	40
2009	20
2008	10
2007	5
2006	4
2005	1

La distribución cronológica de los documentos obtenidos es la siguiente:



Software

I.- R.Temis

En el análisis textual se ha empleado el software **R Text Mining Solution** (R.Temis) (Bouchet-Valat, M. y Bastin, G., 2013). Con R.Temis se pueden realizar operaciones propias de la estadística lexical (medida de las ocurrencias y de las coocurrencias de los términos) y del análisis de datos textuales (clasificación jerárquica ascendente y análisis factorial de correspondencias). R.Temis construye una matriz de documentos (filas) y términos (columnas) y en las celdas de la matriz anota las frecuencias (ocurrencias) de cada término en cada documento.

Se ha utilizado R.Temis para realizar la clasificación de los documentos del corpus de dos maneras diferentes: empleando el texto del campo *index keywords* y empleando el texto del campo *abstract,* para así comprobar los resultados de los dos diferentes análisis.

II.- Software para la técnica de Regresión Lineal Simple

Para la aplicación de la técnica de Regresión Lineal Simple, así como para la validación del modelo y la obtención de distintos gráficos, se ha utilizado el software **SPSS Statistics 20** (IBM).

SPSS es uno de los programas estadísticos más conocidos y utilizados en investigación gracias a su capacidad para trabajar con grandes bases de datos y un sencillo interface para la mayoría de los análisis. Es un software comercial y está compuesto por distintos módulos funcionales que pueden ser comprados e integrados. En nuestro caso, se ha utilizado la versión con licencia que la Universidad de Granada pone a disposición de sus miembros.

Metodología

I. *Hierarchical clustering* y *Correspondence analysis*

Para clasificar los temas de investigación en Investigación traslacional y Medicina Personalizada se ha empleado la técnica del análisis multivariante en los campos *Index keywords* y *Abstract* de los registros de los documentos en la base de datos SCOPUS:

a) Análisis del vocabulario controlado del campo ***Index keywords*** (palabras clave del índice o tesauro), sobre el campo *Index keywords* del archivo CSV de exportación de SCOPUS, y
b) Análisis del vocabulario de lenguaje natural del campo ***Abstract***, sobre un archivo TXT en el que se han copiado los *abstracts* del corpus.

La técnicas multivariantes a emplear para clasificar los temas de investigación han sido la Clasificación ascendente jerárquica (***Hierarchical clustering***) y el Análisis Factorial de correspondencias (***Correspondence analysis***).

El análisis de conglomerados (***cluster***) es una técnica multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencias entre los grupos (Terrádez Gurrea, M.). El dendograma es la representación gráfica utilizada para interpretar el resultado del análisis *cluster*.

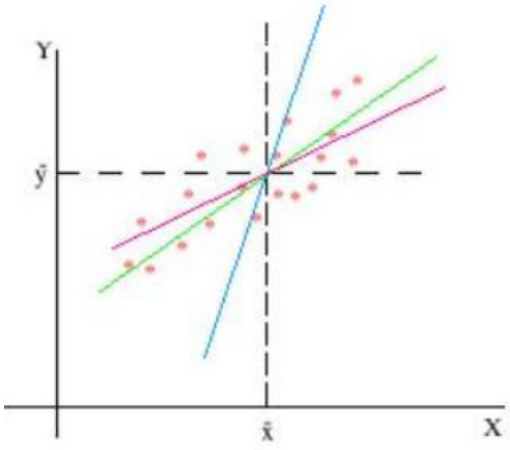
El análisis factorial de correspondencias (AFC) consiste en buscar la mejor representación simultánea de dos conjuntos que constituyen las filas y columnas de una tabla de contingencia. El análisis de correspondencias resume la información contenida en una tabla de contingencia referida a dos conjuntos de grandes dimensiones, teniendo en cuenta el carácter probabilístico de los datos para remediar su heterogeneidad (Navarro Gómez, L., 1983). El análisis factorial de correspondencias permite representar gráficamente las palabras que coocurren en un plano factorial. El análisis de las contribuciones de los términos y la representación gráfica permite encontrar los temas de investigación (Garnier, B).

II. Regresión Lineal Simple

La Regresión Lineal Simple es una técnica estadística que ajusta la relación entre una variable dependiente (Y) y una variable independiente (X) a un modelo lineal (una recta):

$$Y = \beta_0 + \beta_1 X + E$$

El objetivo de la RLS es estimar los parámetros (β_0, β_1) a partir de las observaciones de la variable Y correspondientes a cada una de las x_i, es decir, a partir de los pares de valores (x_1, y_1), …, (x_n, y_n). β_0 es el valor estimado de Y cuando X = 0 y β_1 es la pendiente de la recta.



De todas las posibles rectas que pasan por la nube de puntos formada por los pares (x_n, y_n) se elige aquella (parámetros β_0, β_1) que minimice la suma de los errores o residuos al cuadrado.

El modelo RLS requiere una serie de hipótesis que han de cumplirse para garantizar el correcto funcionamiento de la técnica: linealidad, homogeneidad, homocedasticidad, independencia en las observaciones y normalidad. La fiabilidad del modelo puede estimarse mediante el Coeficiente de Determinación (R^2), el cual mide la proporción de la variabilidad total que viene explicada por el modelo (varianza explicada respecto a varianza total).

Resultados

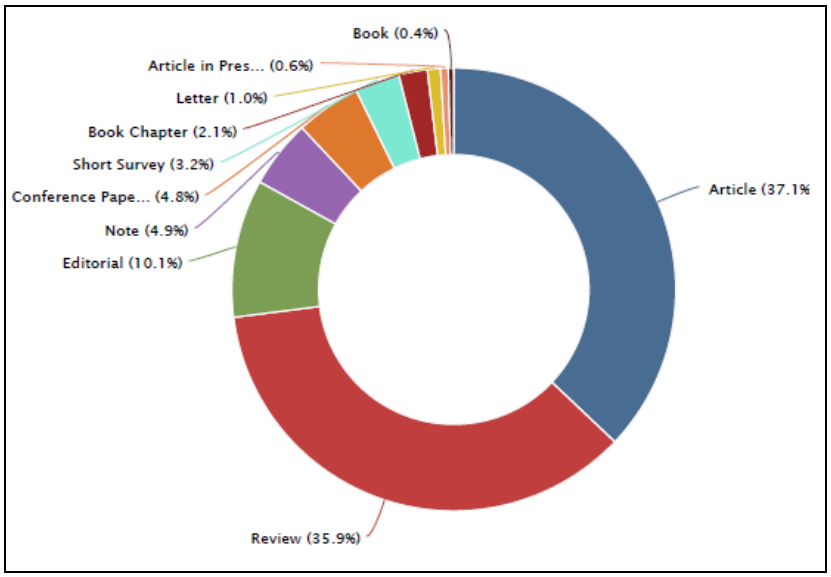
I.- *Hierarchical clustering* y *Correspondence analysis*

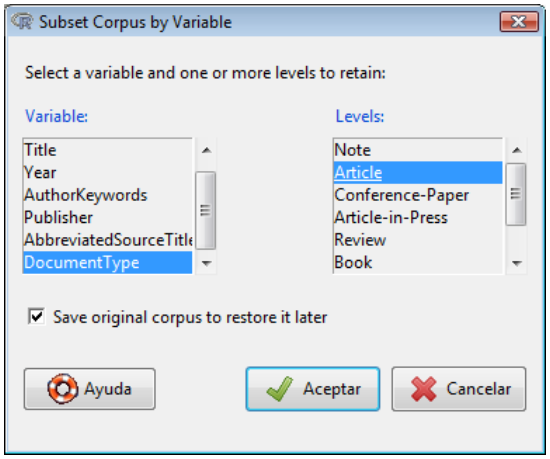
I.1.- Análisis del campo *Index keywords*

El corpus completo descargado de SCOPUS consta de 526 documentos y 4.823 términos.

```
> corpus
<<VCorpus (documents: 526, metadata (corpus/indexed): 3/7)>>

> dtm
<<DocumentTermMatrix (documents: 526, terms: 4823)>>
Non-/sparse entries: 16903/2519995
Sparsity           : 99%
Maximal term length: 136
Weighting          : term frequency (tf)
```

Se ha drealizo el análisis del campo *Index keywords* en el subconjunto de documentos formado por artículos en revistas científicas.



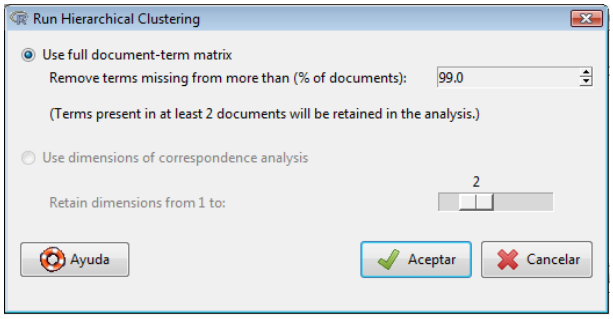


El subconjunto formado por los artículos de revistas científicas consta de 195 documentos.

```
> corpus
<<VCorpus (documents: 195, metadata (corpus/indexed): 3/7)>>

> dtm
<<DocumentTermMatrix (documents: 195, terms: 4823)>>
Non-/sparse entries: 5890/934595
Sparsity           : 99%
Maximal term length: 136
Weighting          : term frequency (tf)
```

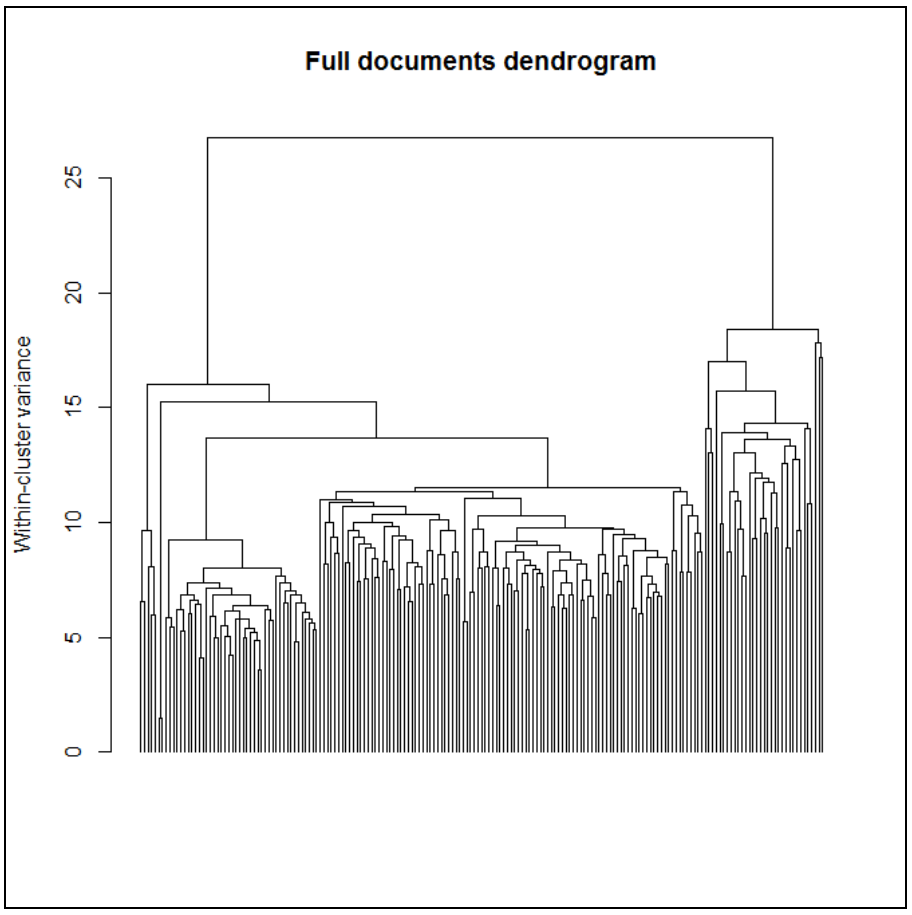
I.1.1.- Clasificación jerárquica ascendente del campo *Index Keywords*

Sobre este conjunto de documentos se ha realizado la técnica de *Hierarchical clustering*. R.Temis realiza el agrupamiento jerárquico ascendente, también llamado agrupamiento jerárquico aglomerativo o acumulativo, según el Método de la varianza mínima de Ward, que utiliza la distancia Chi cuadrado (Bouchet-Valat, M. 2015), procedimiento en el cual, en cada paso, se unen los dos *clusters* para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada cluster, de cada individuo al centroide del *cluster* (Gutiérrez, R., González, A., Torres, F., y Gallardo, J. A. ; 1994) (en la última versión de R.Temis el método de Ward ha sido renombrado Ward D.2). Se han seleccionado los siguientes parámetros:

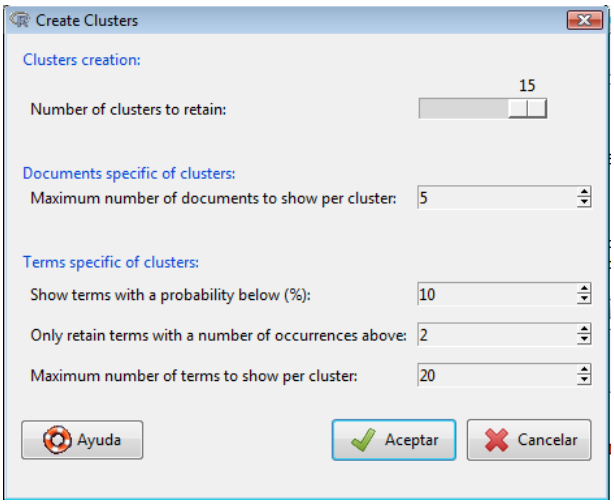


Con los parámetros seleccionados han quedado excluídos del análisis 8 documentos (los documentos número 23, 28, 89, 130, 243, 255, 289 y 311)

Así se ha obtenido el dendograma del conjunto total de documentos:

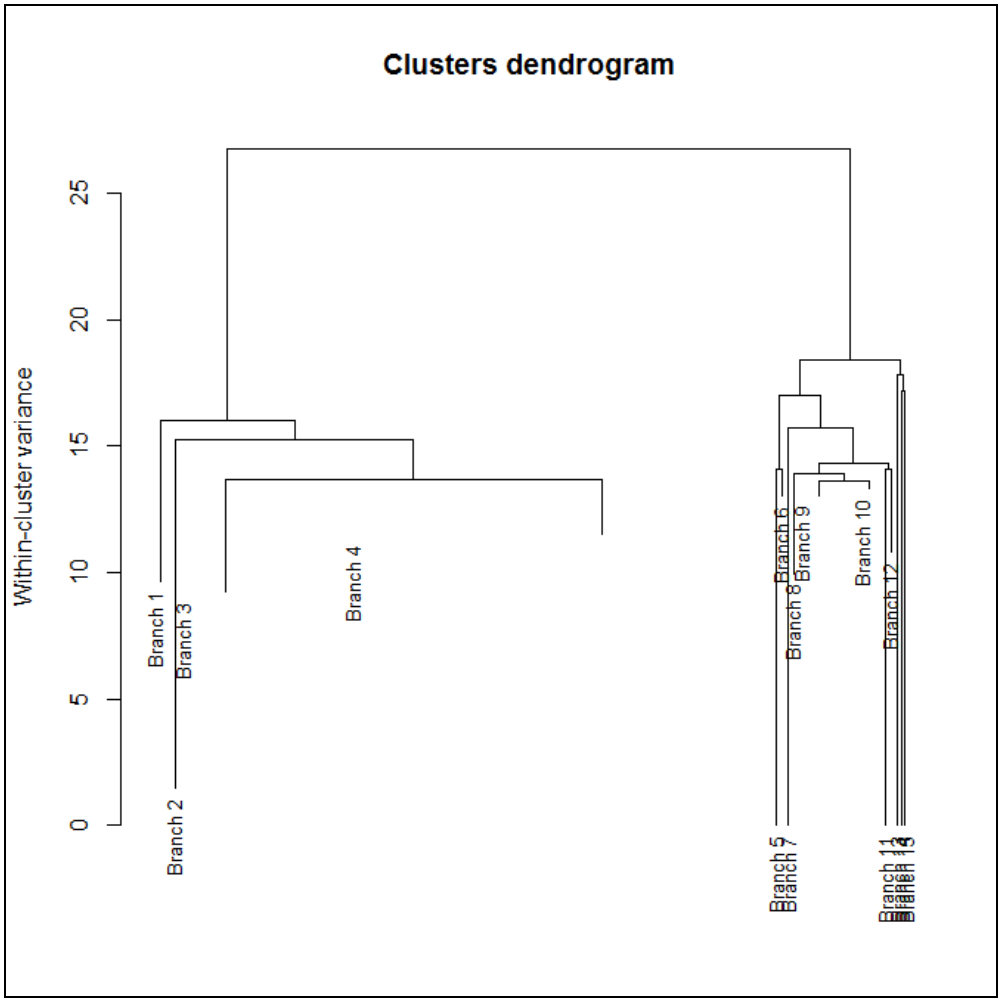


Sobre el total de documentos se han creao 15 agrupaciones de acuerdo con los siguientes parámetros:



Como resultado los 187 documentos (con 767 términos) se han agrupado en 15 *clusters*:

a) Dendrograma de *clusters*:

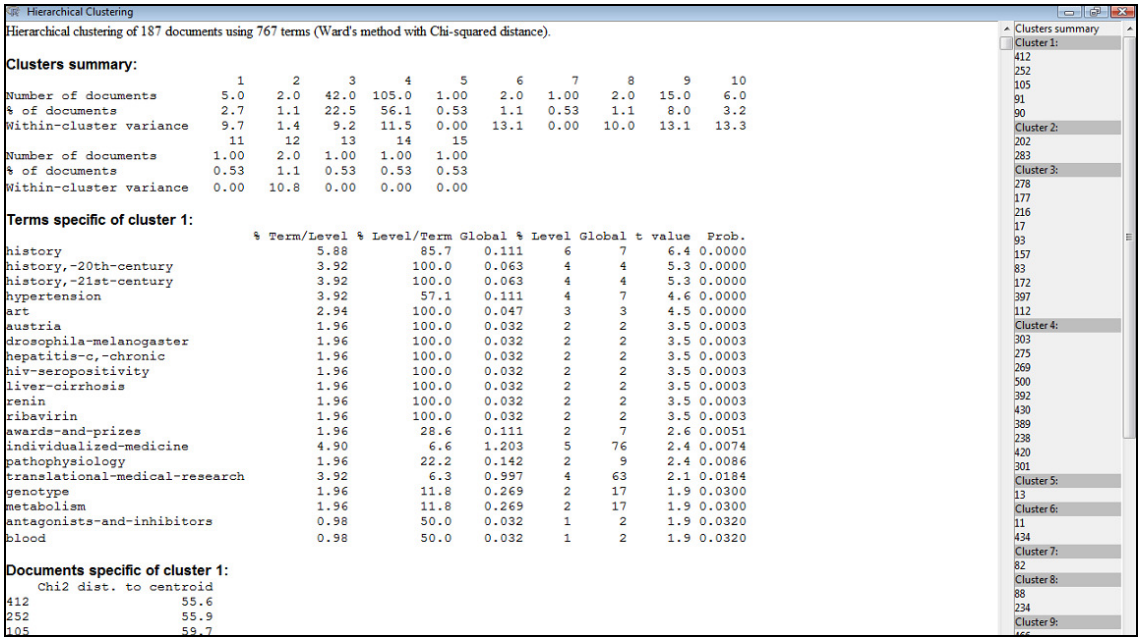


b) Resumen de los *clusters*:

Clusters summary:

	1	2	3	4	5	6	7	8	9	10
Number of documents	5.0	2.0	42.0	105.0	1.00	2.0	1.00	2.0	15.0	6.0
% of documents	2.7	1.1	22.5	56.1	0.53	1.1	0.53	1.1	8.0	3.2
Within-cluster variance	9.7	1.4	9.2	11.5	0.00	13.1	0.00	10.0	13.1	13.3
	11	12	13	14	15					
Number of documents	1.00	2.0	1.00	1.00	1.00					
% of documents	0.53	1.1	0.53	0.53	0.53					
Within-cluster variance	0.00	10.8	0.00	0.00	0.00					

Para mayor información (por ejemplo, documentos que integran los *clusters*) sobre la clasificación jerárquica ascendente del campo *Index Keywords* leer el informe completo en el ANEXO 1



Hierarchical Clustering

Hierarchical clustering of 187 documents using 767 terms (Ward's method with Chi-squared distance).

Clusters summary:

	1	2	3	4	5	6	7	8	9	10
Number of documents	5.0	2.0	42.0	105.0	1.00	2.0	1.00	2.0	15.0	6.0
% of documents	2.7	1.1	22.5	56.1	0.53	1.1	0.53	1.1	8.0	3.2
Within-cluster variance	9.7	1.4	9.2	11.5	0.00	13.1	0.00	10.0	13.1	13.3
	11	12	13	14	15					
Number of documents	1.00	2.0	1.00	1.00	1.00					
% of documents	0.53	1.1	0.53	0.53	0.53					
Within-cluster variance	0.00	10.8	0.00	0.00	0.00					

Terms specific of cluster 1:

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
history	5.88	85.7	0.111	6	7	6.4	0.0000
history,-20th-century	3.92	100.0	0.063	4	4	5.3	0.0000
history,-21st-century	3.92	100.0	0.063	4	4	5.3	0.0000
hypertension	3.92	57.1	0.111	4	7	4.6	0.0000
art	2.94	100.0	0.047	3	3	4.5	0.0000
austria	1.96	100.0	0.032	2	2	3.5	0.0003
drosophila-melanogaster	1.96	100.0	0.032	2	2	3.5	0.0003
hepatitis-c,-chronic	1.96	100.0	0.032	2	2	3.5	0.0003
hiv-seropositivity	1.96	100.0	0.032	2	2	3.5	0.0003
liver-cirrhosis	1.96	100.0	0.032	2	2	3.5	0.0003
renin	1.96	100.0	0.032	2	2	3.5	0.0003
ribavirin	1.96	100.0	0.032	2	2	3.5	0.0003
awards-and-prizes	1.96	28.6	0.111	2	7	2.6	0.0051
individualized-medicine	4.90	6.6	1.203	5	76	2.4	0.0074
pathophysiology	1.96	22.2	0.142	2	9	2.4	0.0086
translational-medical-research	3.92	6.3	0.997	4	63	2.1	0.0184
genotype	1.96	11.8	0.269	2	17	1.9	0.0300
metabolism	1.96	11.8	0.269	2	17	1.9	0.0300
antagonists-and-inhibitors	0.98	50.0	0.032	1	2	1.9	0.0320
blood	0.98	50.0	0.032	1	2	1.9	0.0320

Documents specific of cluster 1:

	Chi2 dist. to centroid
412	55.6
252	55.9
105	59.7

Clusters summary
Cluster 1:
412
252
105
91
90
Cluster 2:
202
283
Cluster 3:
278
177
216
17
93
157
83
172
397
112
Cluster 4:
303
275
269
500
392
430
389
238
420
301
Cluster 5:
13
Cluster 6:
11
434
Cluster 7:
82
Cluster 8:
88
234
Cluster 9:

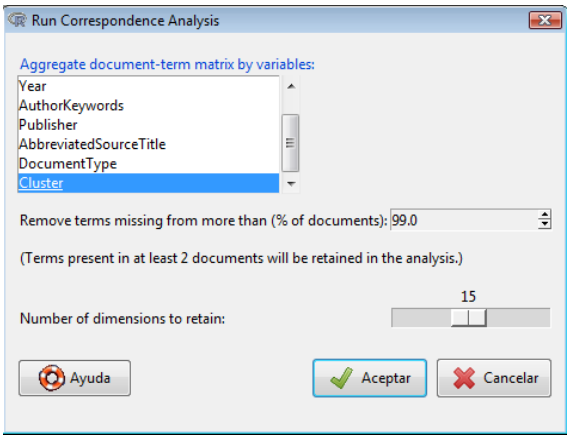
"% Term/Level": the percent of the term's occurrences in all terms occurrences in the level.
"% Level/Term": the percent of the term's occurrences that appear in the level (rather than in other levels).
"Global %": the percent of the term's occurrences in all terms occurrences in the corpus.
"Level": the number of occurrences of the term in the level ("internal").
"Global": the number of occurrences of the term in the corpus.
"t value": the quantile of a normal distribution corresponding the probability "Prob.".
"Prob.": the probability of observing such an extreme (high or low) number of occurrences of the term in the level, under an hypergeometric distribution.

I.1.2.- Análisis factorial de correspondencias del campo *Index keywords*

Se ha realizado el Análisis factorial de correspondencias sobre la matriz documentos-términos y las clasificaciones obtenidas en el agrupamiento jerárquico ascendente para así estructurar el conjunto de términos en función de las mismas.

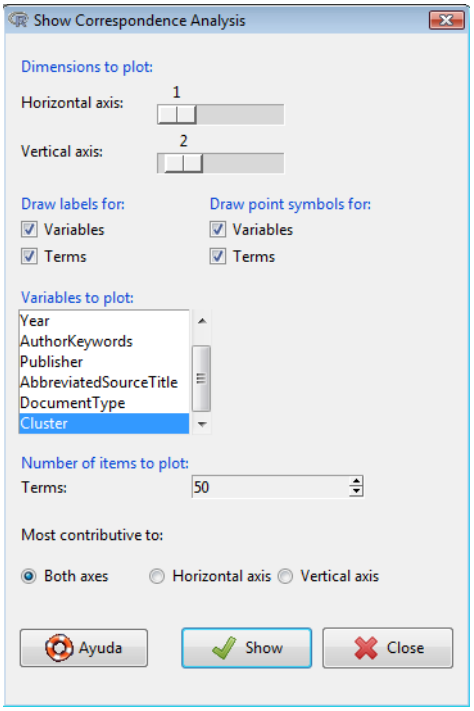
Si no se selecciona ninguna variable de la lista (por defecto en R.Temis), el análisis de correspondencias se ejecuta en la matriz completa de documentos/términos. Si se eligen una o más variables, el análisis de correspondencias se realizará sobre una tabla cuyas filas

corresponden a los niveles de la variable, y donde cada celda contiene la suma de ocurrencias de un término dado en todos los documentos del nivel.



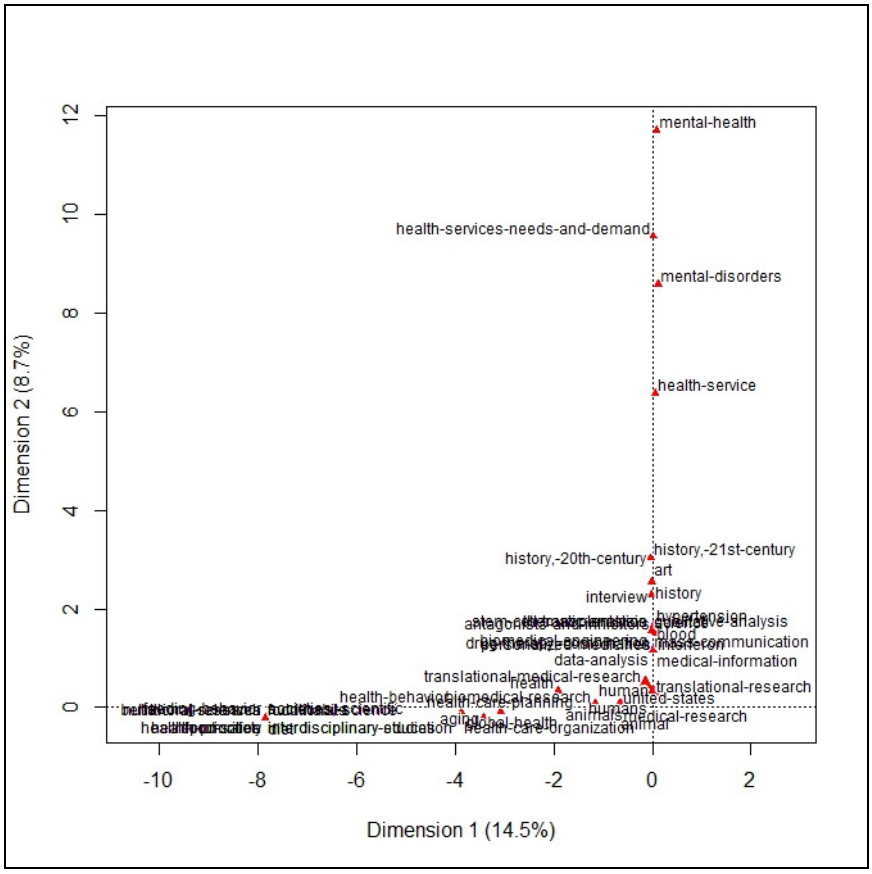
8 documentos han sido eliminados del análisis por no cumplir los parámetros seleccionados: 23, 28, 89, 130, 243, 255, 289 y 311.

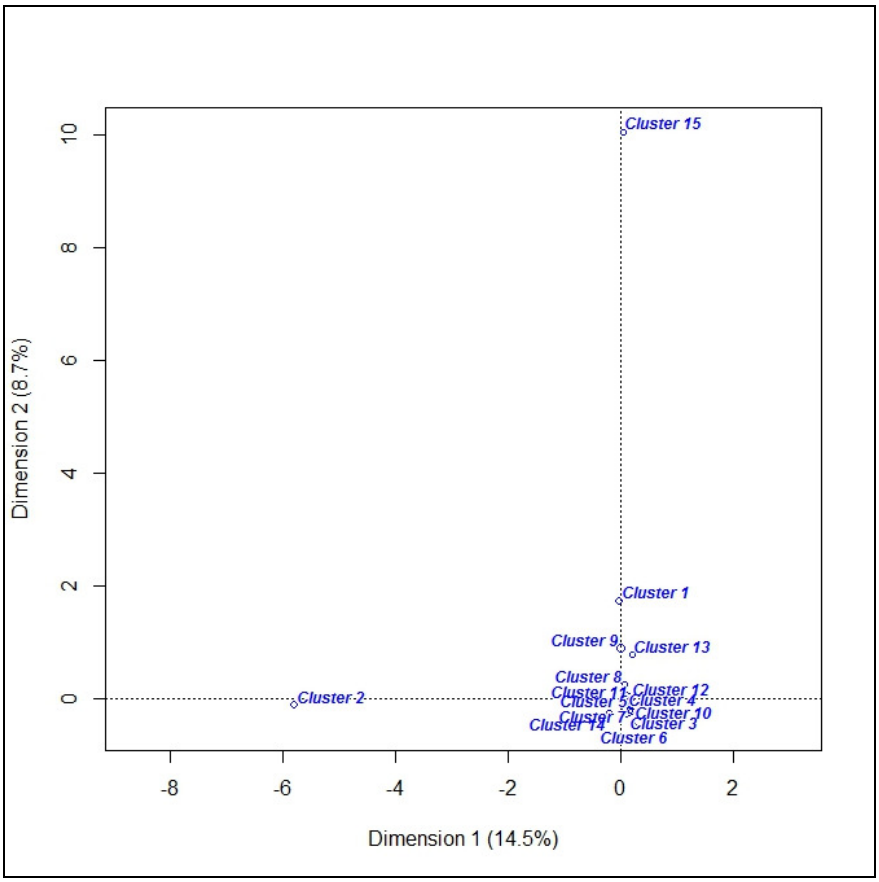
Como consecuencia se ha realizado el análisis de correspondencias empleando 187 documentos (con 767 términos).

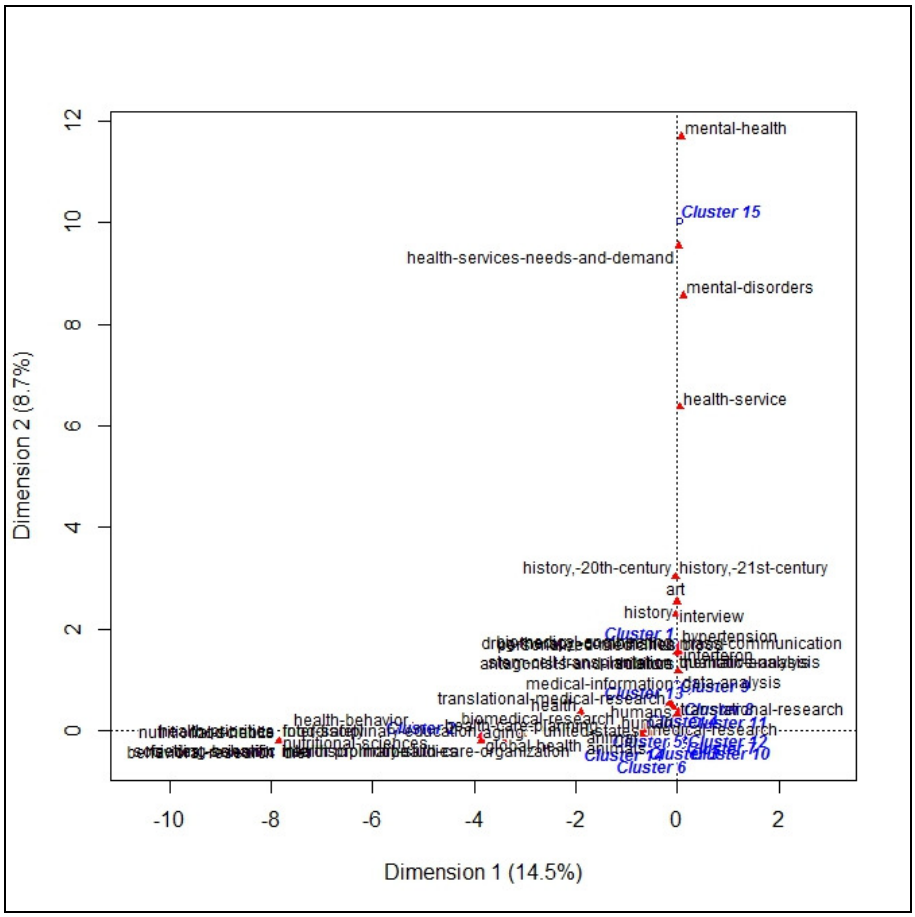


Para mayor información sobre el análisis factorial de correspondencias del campo *Index Keywords* leer el informe completo en el ANEXO 3.

a) Gráficos del análisis de correspondencias:







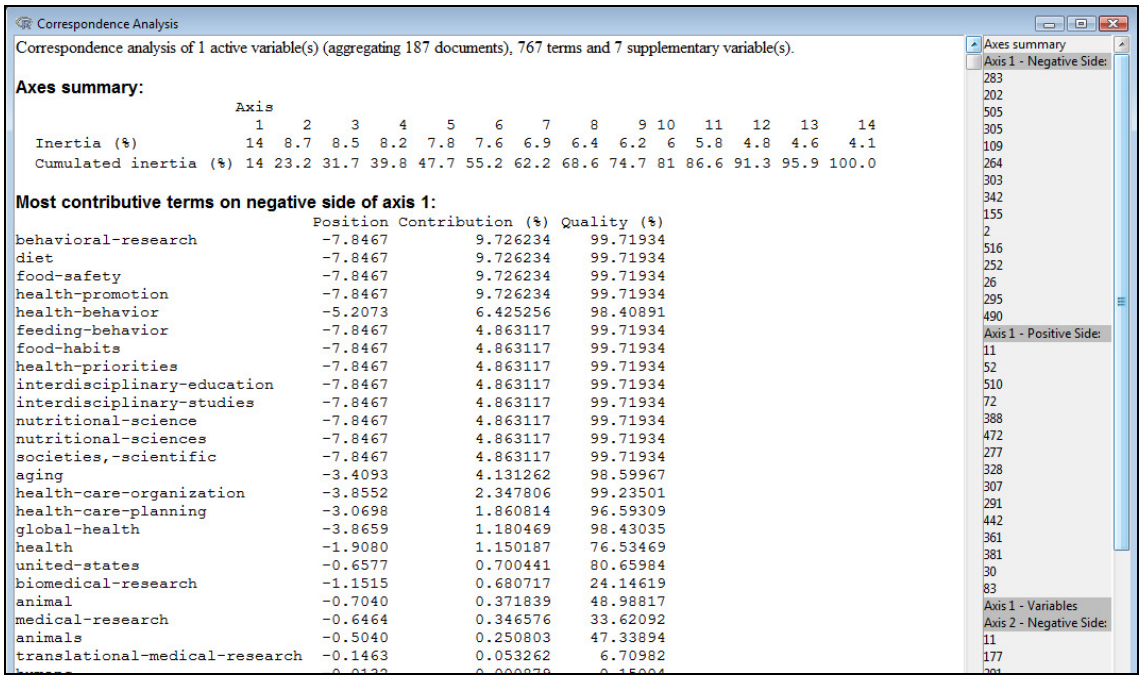
b) Interpretación de los gráficos:

La ventana de texto muestra, en primer lugar el porcentage de la inercia total de cada eje [la inercia total es una medida similar a la variación total en el caso de los componentes principales, mide el grado total de dependencia existente entre las variables (Salvador Figueras, M. 2003)], y a continuación su inercia acumulada. En el resto de la ventana de texto sólo aparecen los datos de los ejes seleccionados (horizontal y vertical): los elementos activos más contributivos a la dimensión (eje) elegida, junto con su posición (*Position*), la contribución del elemento a la inercia de la dimensión, en %, (*Contribution*) y la contribución de la dimensión a la inercia del elemento, en %, (*Quality*) (Bouchet-Valat, M. 2015).

Si se busca determinar un sentido de causalidad en la relación *Contribution* – *Quality* [correlación con el factor o dimensión de un elemento (Navarro Gómez, L. 1983)], hay que comparar la serie de los valores de *Quality* con la de los valores de *Contribution*. Pero como una fuerte *Contribution* no implica una *Quality* elevada, esto hace que la utilización de las *Qualities* sea delicada y se hace necesario el análisis de *Position*, *Contribution* y *Quality* para poder dar una interpretación correcta y precisa a los factores o dimensiones (Navarro Gómez, L. 1983).

Se observa que las dimensiones o factores obtenidos más representativos del conjunto son el 1, 2 y 3, con una inercia entre el 8,5 y el 14 %, las dos primeras dimensiones sólo recogen el 25% de la inercia acumulada.

Se observa que la primera dimensión y la segunda dimensión suman una inercia acumulada de del 23,2 %.



Los términos más contributivos en la zona negativa del eje 1 son:

	Position	Contribution (%)	Quality (%)
behavioral-research	-7.8467	9.726234	99.71934
diet	-7.8467	9.726234	99.71934
food-safety	-7.8467	9.726234	99.71934
health-promotion	-7.8467	9.726234	99.71934
health-behavior	-5.2073	6.425256	98.40891
feeding-behavior	-7.8467	4.863117	99.71934
food-habits	-7.8467	4.863117	99.71934
health-priorities	-7.8467	4.863117	99.71934
interdisciplinary-education	-7.8467	4.863117	99.71934
interdisciplinary-studies	-7.8467	4.863117	99.71934
nutritional-science	-7.8467	4.863117	99.71934
nutritional-sciences	-7.8467	4.863117	99.71934
societies,-scientific	-7.8467	4.863117	99.71934
aging	-3.4093	4.131262	98.59967
health-care-organization	-3.8552	2.347806	99.23501
health-care-planning	-3.0698	1.860814	96.59309
global-health	-3.8659	1.180469	98.43035
health	-1.9080	1.150187	76.53469
united-states	-0.6577	0.700441	80.65984
biomedical-research	-1.1515	0.680717	24.14619
animal	-0.7040	0.371839	48.98817
medical-research	-0.6464	0.346576	33.62092
animals	-0.5040	0.250803	47.33894
translational-medical-research	-0.1463	0.053262	6.70982
humans	-0.0132	0.000879	0.15004
history,-20th-century	-0.0351	0.000195	0.00211
history,-21st-century	-0.0351	0.000195	0.00211
art	-0.0351	0.000146	0.00211

interview	-0.0251	0.000050	0.00326
biomedical-engineering	-0.0150	0.000044	0.00110
stem-cell-transplantation	-0.0150	0.000044	0.00110
science	-0.0150	0.000027	0.00110
mass-communication	-0.0150	0.000018	0.00110
qualitative-analysis	-0.0150	0.000018	0.00110
thematic-analysis	-0.0150	0.000018	0.00110
history	-0.0076	0.000016	0.00013

Los grupos activos en la zona negativa del eje 1 son:

	Position	Contribution (%)	Quality (%)
Cluster 2	-5.795	98.1300	99.8504
Cluster 14	-0.217	0.0130	0.0321
Cluster 1	-0.026	0.0021	0.0036
Cluster 9	-0.011	0.0010	0.0022

Los documentos más alejados en la zona negativa del eje 1 son:

	Position	Quality (%)
283	-5.87	98.362
202	-5.72	98.320
505	-0.36	0.524
305	-0.26	0.315
109	-0.22	0.032
264	-0.22	0.108
303	-0.21	0.190
342	-0.20	0.134
155	-0.20	0.068
2	-0.19	0.097
516	-0.18	0.104
252	-0.18	0.100
26	-0.13	0.036
295	-0.12	0.029
490	-0.12	0.082

Los términos más contributivos en la zona positiva del eje 1 son:

	Position	Contribution (%)	Quality (%)
human	0.020	0.002701	0.35584
translational-research	0.020	0.002700	0.21812
mental-disorders	0.110	0.001905	0.01042
mental-health	0.079	0.000748	0.00307
personalized-medicines	0.078	0.000722	0.00741
hypertension	0.035	0.000344	0.00667
health-service	0.054	0.000344	0.00545
antagonists-and-inhibitors	0.040	0.000125	0.01103
blood	0.040	0.000125	0.01103
drug-therapy,-combination	0.040	0.000125	0.01103
interferon	0.040	0.000125	0.01103
data-analysis	0.017	0.000048	0.00273
medical-information	0.017	0.000048	0.00273

health-services-needs-and-demand	0.023	0.000043	0.00045

Los grupos activos en la zona positiva del eje 1 son:

	Position	Contribution (%)	Quality (%)
Cluster 3	0.117	0.93486	2.9127
Cluster 4	0.085	0.66149	2.8972
Cluster 10	0.210	0.16190	0.2984
Cluster 6	0.215	0.03301	0.0656
Cluster 12	0.169	0.02603	0.0573
Cluster 11	0.156	0.01054	0.0248
Cluster 13	0.195	0.01052	0.0232
Cluster 5	0.149	0.00699	0.0200
Cluster 7	0.132	0.00415	0.0146
Cluster 8	0.070	0.00353	0.0104
Cluster 15	0.046	0.00082	0.0014

Los documentos más alejados en la zona positiva del eje 1 son:

	Position	Quality (%)
11	0.25	0.054
52	0.23	0.100
510	0.23	0.098
72	0.21	0.068
388	0.21	0.059
472	0.20	0.023
277	0.18	0.038
328	0.18	0.036
307	0.18	0.041
291	0.18	0.168
442	0.18	0.209
361	0.18	0.155
381	0.17	0.160
30	0.17	0.098
83	0.17	0.170

Los términos más contributivos en la zona negativa del eje 2 son :

	Position	Contribution (%)	Quality (%)
aging	-0.228	0.0306	0.440
animals	-0.120	0.0238	2.702
animal	-0.117	0.0171	1.357
behavioral-research	-0.213	0.0119	0.073
diet	-0.213	0.0119	0.073
food-safety	-0.213	0.0119	0.073
health-promotion	-0.213	0.0119	0.073
health-care-organization	-0.206	0.0111	0.283
feeding-behavior	-0.213	0.0059	0.073
food-habits	-0.213	0.0059	0.073
health-priorities	-0.213	0.0059	0.073
interdisciplinary-education	-0.213	0.0059	0.073

interdisciplinary-studies	-0.213	0.0059	0.073
nutritional-science	-0.213	0.0059	0.073
nutritional-sciences	-0.213	0.0059	0.073
societies,-scientific	-0.213	0.0059	0.073
health-care-planning	-0.082	0.0022	0.068
global-health	-0.103	0.0014	0.070

Los grupos activos en la zona negativa del eje 2 son :

	Position	Contribution (%)	Quality (%)
Cluster 3	-0.231	6.11387	11.46454
Cluster 10	-0.226	0.31074	0.34474
Cluster 6	-0.487	0.27980	0.33445
Cluster 2	-0.122	0.07202	0.04411
Cluster 14	-0.263	0.03180	0.04719
Cluster 5	-0.197	0.02030	0.03498
Cluster 7	-0.125	0.00613	0.01303
Cluster 11	-0.028	0.00057	0.00081

Los documentos más alejados en la zona negativa del eje 2 son:

	Position	Quality (%)
11	-0.65	0.37
177	-0.41	1.74
291	-0.40	0.83
172	-0.38	0.83
381	-0.34	0.60
186	-0.34	0.52
93	-0.34	0.90
99	-0.33	0.40
30	-0.32	0.34
52	-0.32	0.19
72	-0.32	0.15
112	-0.31	0.53
398	-0.31	0.66
216	-0.30	0.77
442	-0.29	0.57

Los términos más contributivos en la zona positiva del eje 2 son :

	Position	Contribution (%)	Quality (%)
mental-health	11.694	26.91823	66.5155
mental-disorders	8.567	19.26363	63.4106
health-services-needs-and-demand	9.544	11.95483	75.6435
health-service	6.365	7.97470	76.1649
history	2.531	2.94307	14.9747
history,-20th-century	3.020	2.39454	15.5879
history,-21st-century	3.020	2.39454	15.5879
art	3.020	1.79591	15.5879
hypertension	1.670	1.28093	14.9496
translational-medical-research	0.474	0.92951	70.4754
biomedical-engineering	1.551	0.78927	11.7944

stem-cell-transplantation	1.551	0.78927	11.7944
translational-research	0.259	0.73990	35.9701
interview	2.286	0.68565	27.1677
science	1.551	0.47356	11.7944
personalized-medicines	1.487	0.43504	2.6883
human	0.193	0.39770	31.5328
data-analysis	1.165	0.35605	12.1553
medical-information	1.165	0.35605	12.1553
humans	0.207	0.35560	36.5179
mass-communication	1.551	0.31571	11.7944
qualitative-analysis	1.551	0.31571	11.7944
thematic-analysis	1.551	0.31571	11.7944
antagonists-and-inhibitors	1.513	0.30046	15.9218
blood	1.513	0.30046	15.9218
drug-therapy,-combination	1.513	0.30046	15.9218
interferon	1.513	0.30046	15.9218
health	0.337	0.05979	2.3945
biomedical-research	0.055	0.00261	0.0556
united-states	0.029	0.00233	0.1612
medical-research	0.040	0.00220	0.1285
health-behavior	0.050	0.00096	0.0089

<u>Los grupos activos en la zona positiva del eje 2 son :</u>

	Position	Contribution (%)	Quality (%)
Cluster 15	10.0474	66.24270	68.35221
Cluster 1	1.7304	15.32606	16.07683
Cluster 9	0.8886	11.24314	14.50713
Cluster 13	0.7779	0.27794	0.36899
Cluster 8	0.2488	0.07314	0.13014
Cluster 4	0.0033	0.00167	0.00441
Cluster 12	0.0088	0.00012	0.00015

<u>Los documentos más alejados en la zona positiva del eje 2 son :</u>

	Position	Quality (%)
191	10.05	68.35
90	2.19	9.72
466	1.99	8.07
91	1.94	6.28
105	1.84	6.77
252	1.53	7.66
176	1.27	2.72
412	1.27	6.66
340	1.02	1.82
408	0.97	1.36
139	0.92	1.33
436	0.82	1.08
472	0.78	0.37
292	0.74	1.30
390	0.70	0.86

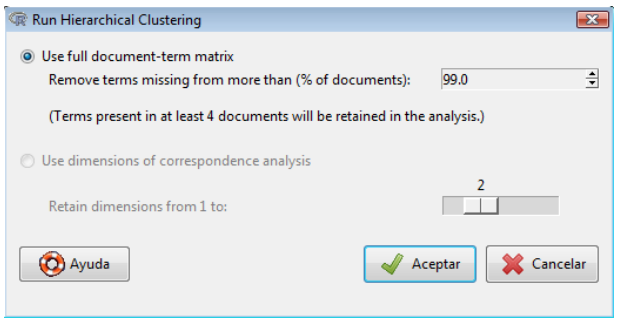
I.2- Análisis del campo *Abstract*.

El subconjunto del campo *abstract* consta de 396 documentos, con 7145 términos.

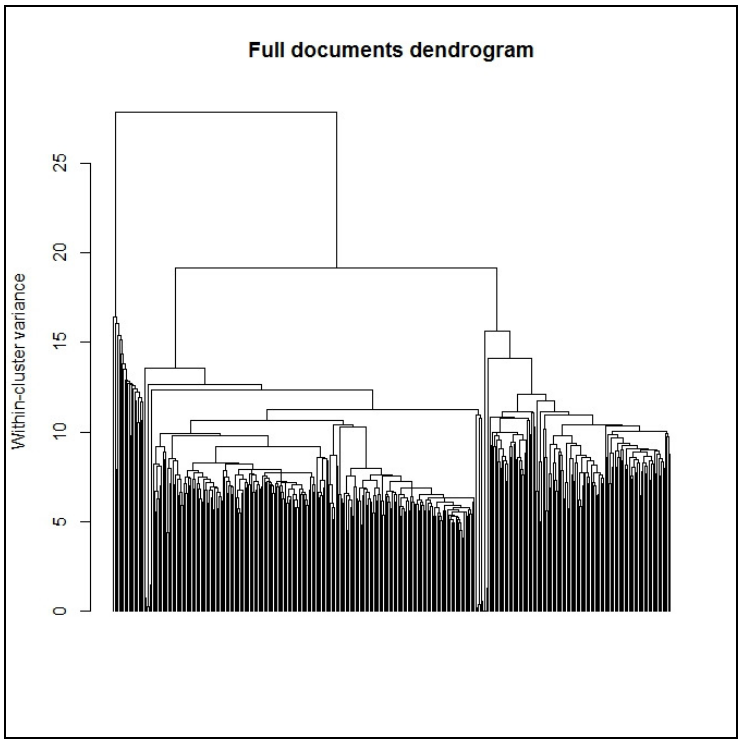
```
> corpus
<<VCorpus (documents: 396, metadata (corpus/indexed): 4/2)>>

> dtm
<<DocumentTermMatrix (documents: 396, terms: 7145)>>
Non-/sparse entries: 37964/2791456
Sparsity           : 99%
Maximal term length: 23
Weighting          : term frequency (tf)
```

I.2.1.- Clasificación jerárquica ascendente del campo *Abstract*

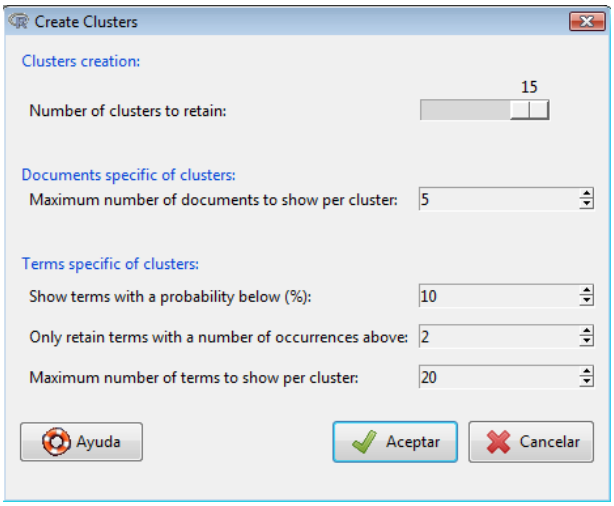
Sobre este conjunto de documentos se ha realizado la técnica *Hierarchical clustering* con los siguientes parámetros.



Como resultado se ha obtenido el dendrograma del conjunto total de documentos:

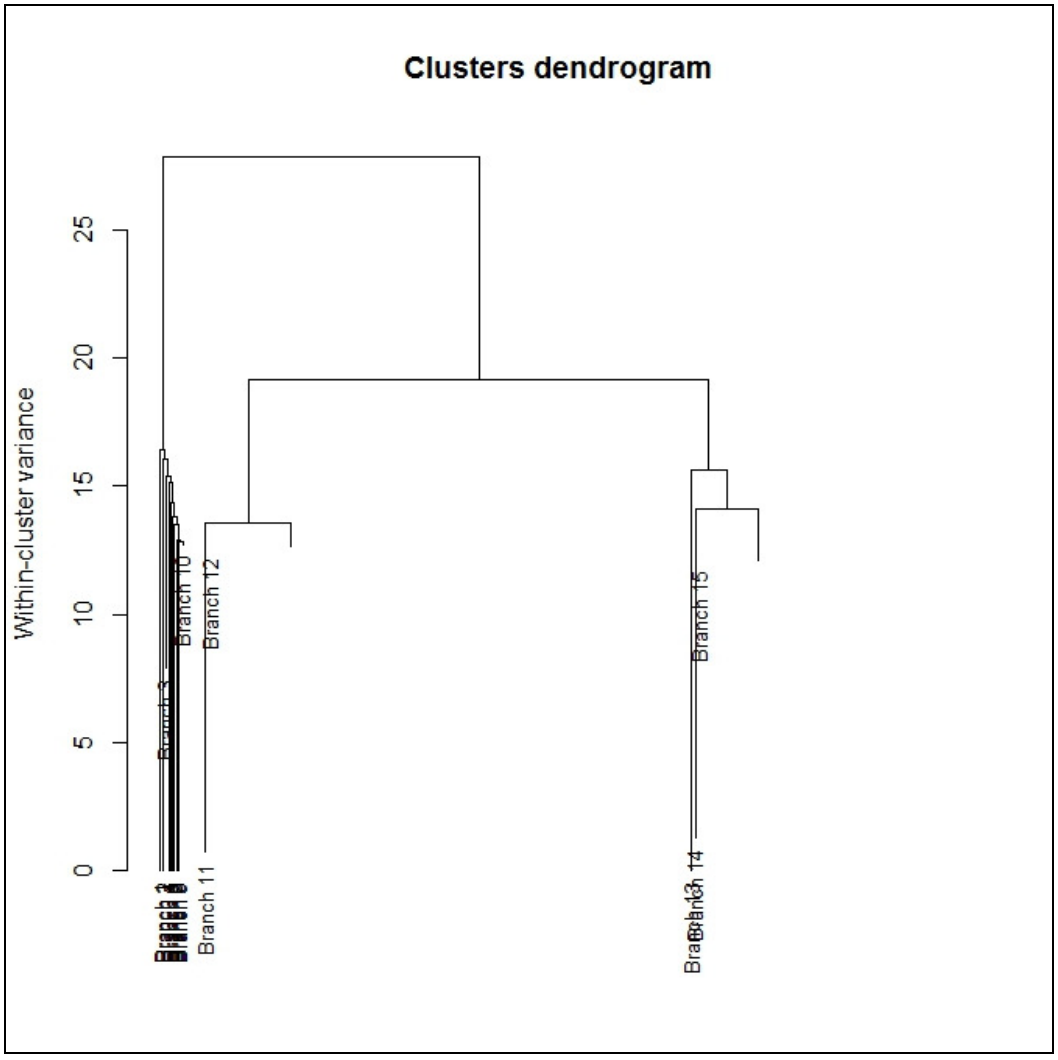


Sobre el total de documentos se han creao 15 agrupaciones de acuerdo con los siguientes parámetros:



Y se ha obtenido el agrupamiento siguiente de los 396 documentos (con 2002 términos):

a) Dendrograma de los *clusters*:

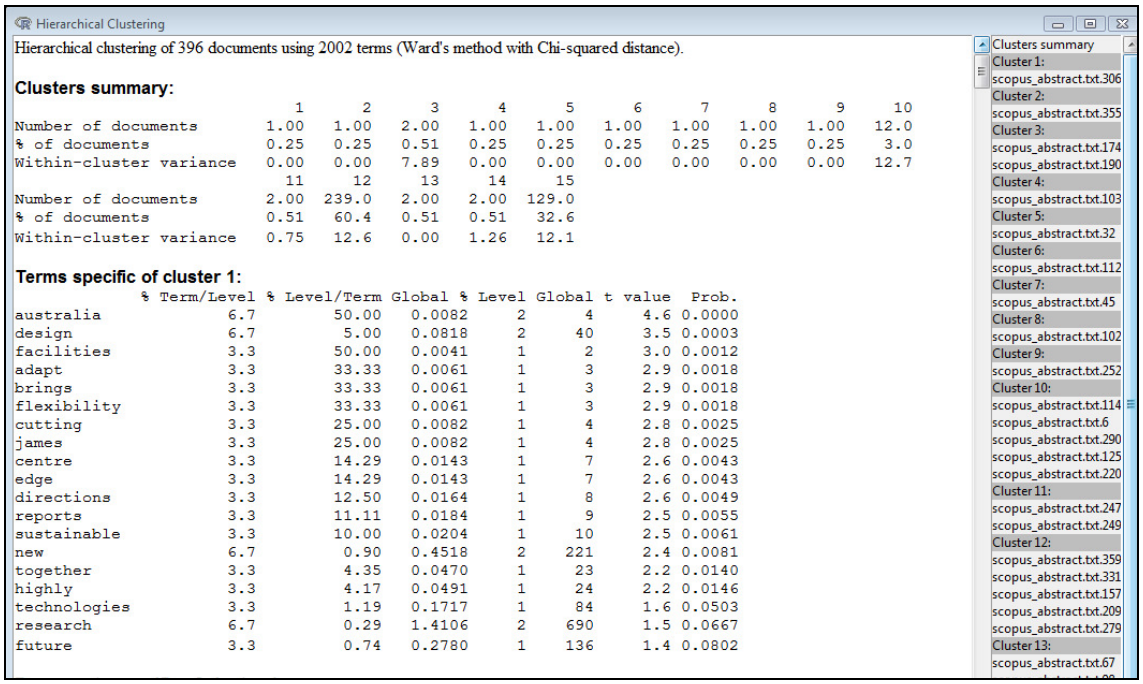


b) Resumen de los *clusters*:

Clusters summary:

	1	2	3	4	5	6	7	8	9	10
Number of documents	1.00	1.00	2.00	1.00	1.00	1.00	1.00	1.00	1.00	12.0
% of documents	0.25	0.25	0.51	0.25	0.25	0.25	0.25	0.25	0.25	3.0
Within-cluster variance	0.00	0.00	7.89	0.00	0.00	0.00	0.00	0.00	0.00	12.7
	11	12	13	14	15					
Number of documents	2.00	239.0	2.00	2.00	129.0					
% of documents	0.51	60.4	0.51	0.51	32.6					
Within-cluster variance	0.75	12.6	0.00	1.26	12.1					

Para mayor información sobre el agrupamiento jerárquico ascendente del campo *Asbtract* leer el informe completo en el ANEXO 2.



	1	2	3	4	5	6	7	8	9	10
Number of documents	1.00	1.00	2.00	1.00	1.00	1.00	1.00	1.00	1.00	12.0
% of documents	0.25	0.25	0.51	0.25	0.25	0.25	0.25	0.25	0.25	3.0
Within-cluster variance	0.00	0.00	7.89	0.00	0.00	0.00	0.00	0.00	0.00	12.7
	11	12	13	14	15					
Number of documents	2.00	239.0	2.00	2.00	129.0					
% of documents	0.51	60.4	0.51	0.51	32.6					
Within-cluster variance	0.75	12.6	0.00	1.26	12.1					

Terms specific of cluster 1:

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
australia	6.7	50.00	0.0082	2	4	4.6	0.0000
design	6.7	5.00	0.0818	2	40	3.5	0.0003
facilities	3.3	50.00	0.0041	1	2	3.0	0.0012
adapt	3.3	33.33	0.0061	1	3	2.9	0.0018
brings	3.3	33.33	0.0061	1	3	2.9	0.0018
flexibility	3.3	33.33	0.0061	1	3	2.9	0.0018
cutting	3.3	25.00	0.0082	1	4	2.8	0.0025
james	3.3	25.00	0.0082	1	4	2.8	0.0025
centre	3.3	14.29	0.0143	1	7	2.6	0.0043
edge	3.3	14.29	0.0143	1	7	2.6	0.0043
directions	3.3	12.50	0.0164	1	8	2.6	0.0049
reports	3.3	11.11	0.0184	1	9	2.5	0.0055
sustainable	3.3	10.00	0.0204	1	10	2.5	0.0061
new	6.7	0.90	0.4518	2	221	2.4	0.0081
together	3.3	4.35	0.0470	1	23	2.2	0.0140
highly	3.3	4.17	0.0491	1	24	2.2	0.0146
technologies	3.3	1.19	0.1717	1	84	1.6	0.0503
research	6.7	0.29	1.4106	2	690	1.5	0.0667
future	3.3	0.74	0.2780	1	136	1.4	0.0802

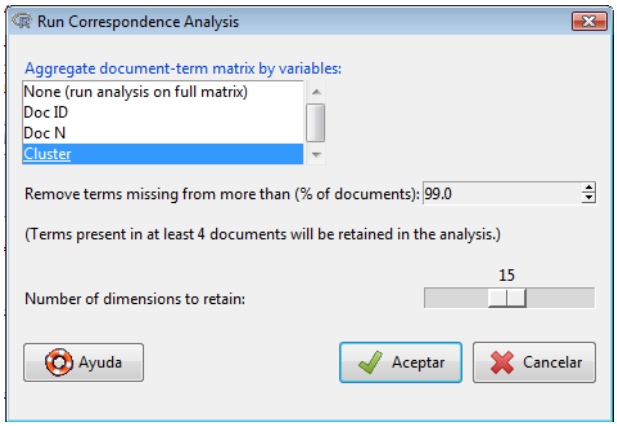
"% Term/Level": the percent of the term's occurrences in all terms occurrences in the level.
"% Level/Term": the percent of the term's occurrences that appear in the level (rather than in other levels).
"Global %": the percent of the term's occurrences in all terms occurrences in the corpus.
"Level": the number of occurrences of the term in the level ("internal").
"Global": the number of occurrences of the term in the corpus.
"t value": the quantile of a normal distribution corresponding the probability "Prob.".
"Prob.": the probability of observing such an extreme (high or low) number of occurrences of the term in the level, under an hypergeometric distribution.

I.2.2.- Análisis factorial de correspondencias del campo *Abstract*

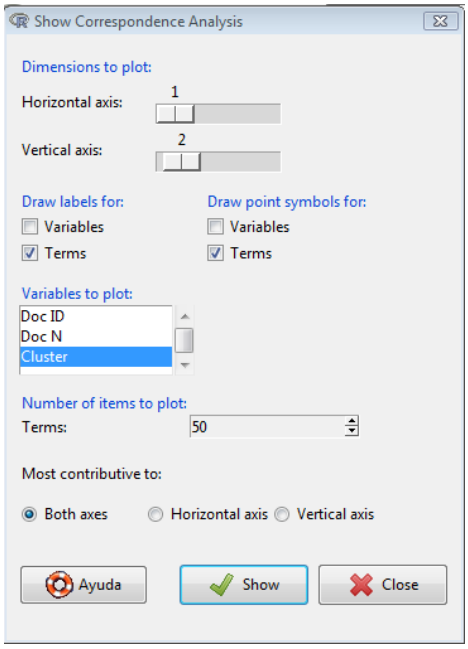
Se ha realizado elanálisis de correspondencias de la variable activa *cluster* agregando 396 documentos (2002 terminos).

Si no se selecciona ninguna variable de la lista (por defecto en R.Temis), el análisis de correspondencias se ejecuta en la matriz completa de documentos/términos. Si se eligen una o más variables, el análisis de correspondencias se realizará sobre una tabla cuyas filas corresponden a los niveles de la variable, y donde cada celda contiene la suma de ocurrencias de un término dado en todos los documentos del nivel.

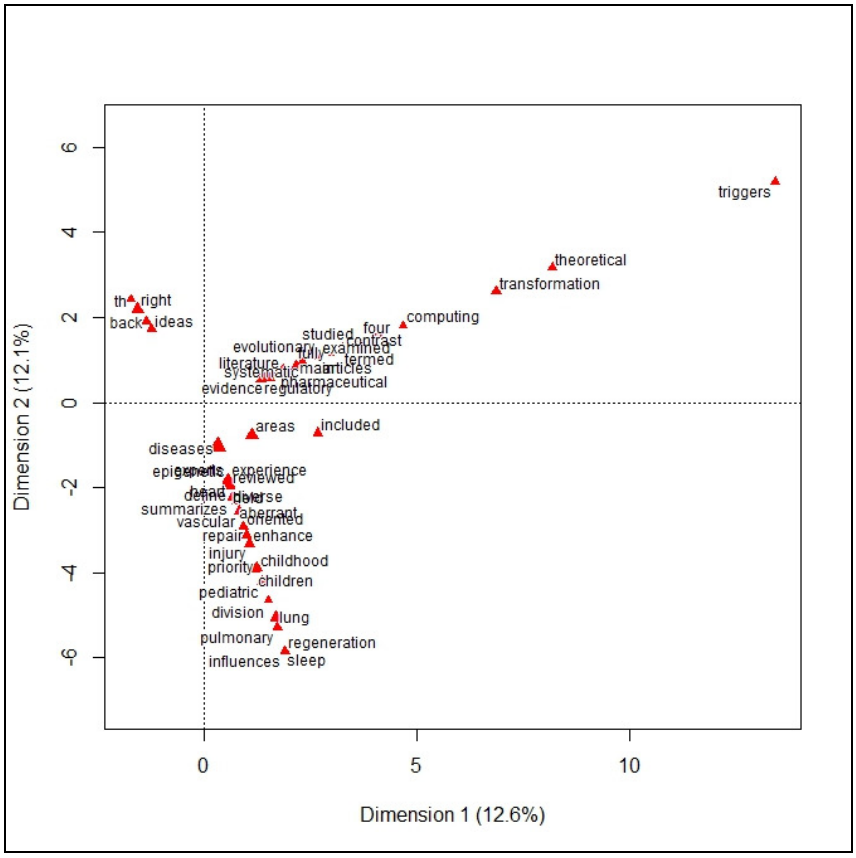
Los parámetros del análisis han sido:

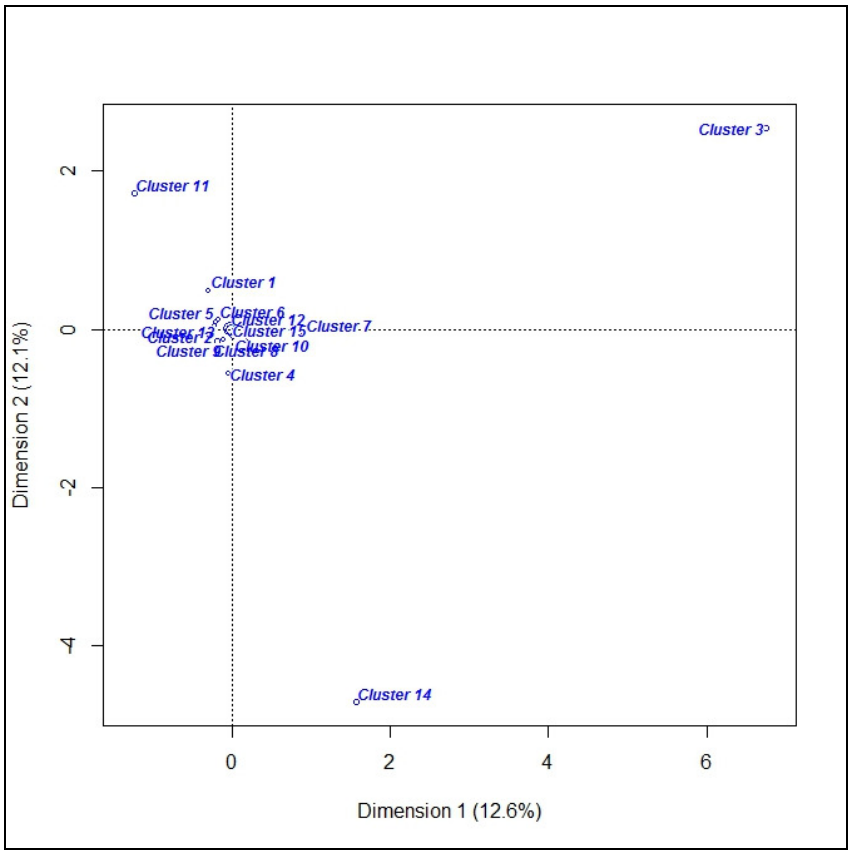


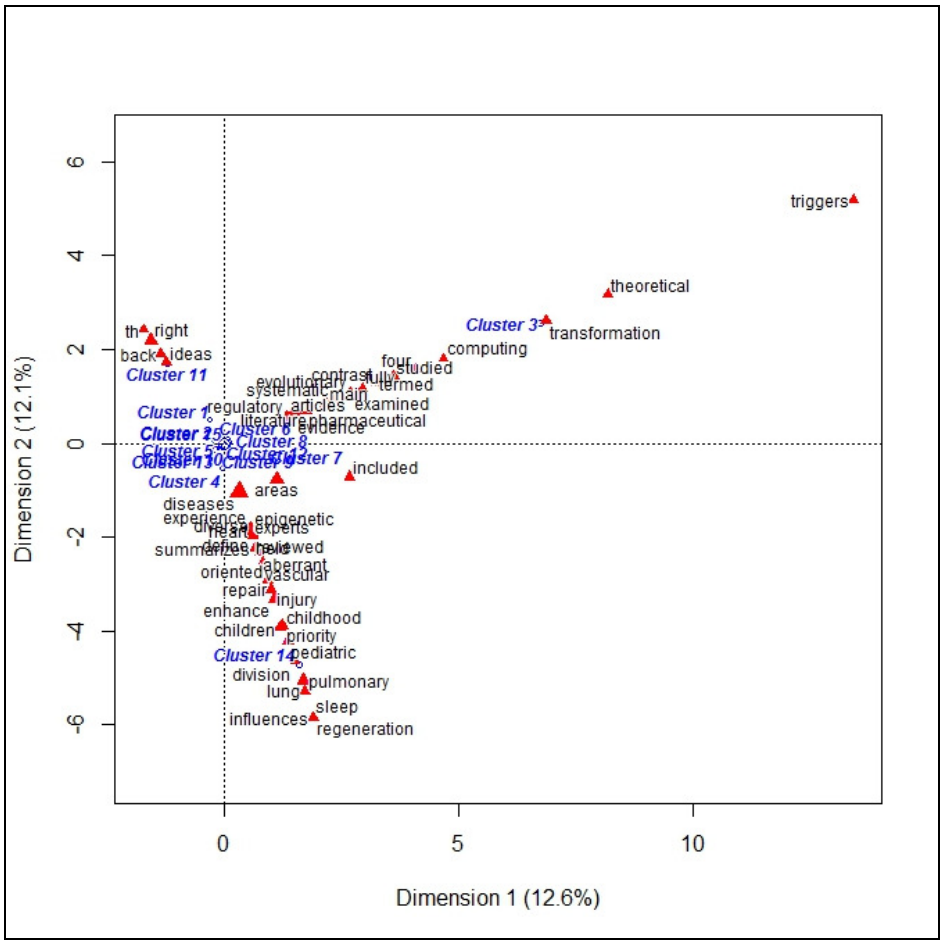
Y los parámetros de representación del análisis en los planos factoriales han sido:



a) Gráficos del análisis de correspondencias:





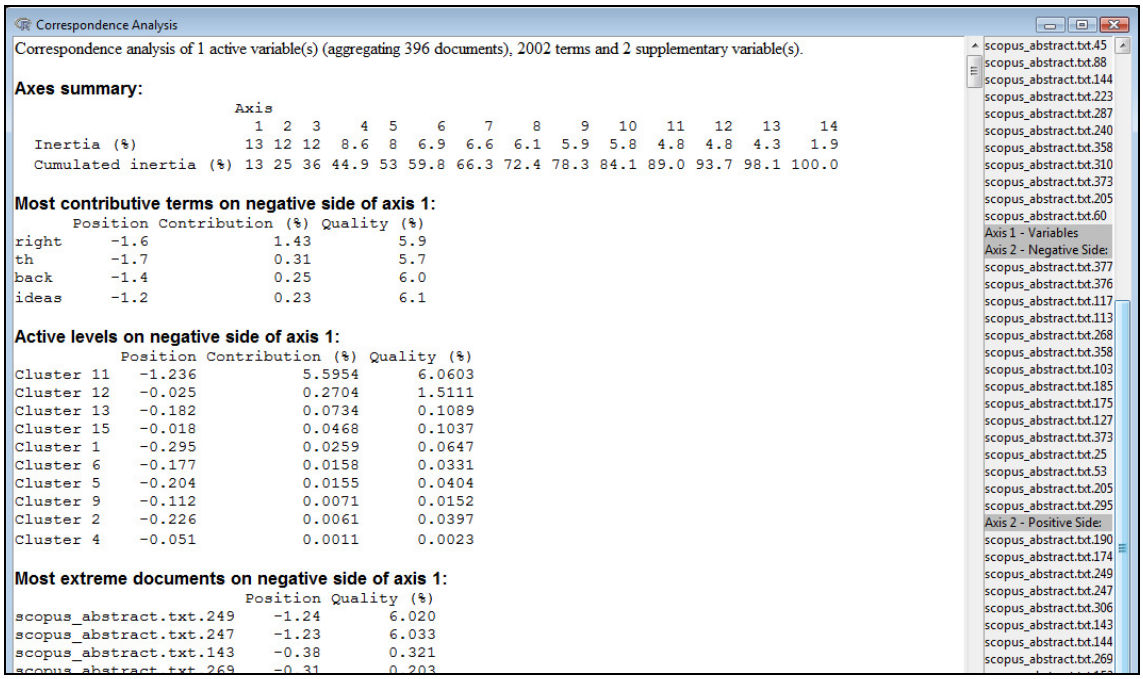


b) Interpretación de los gráficos:

La ventana de texto muestra, en primer lugar el porcentage de la inercia total de cada eje [la inercia total es una medida similar a la variación total en el caso de los componentes principales, mide el grado total de dependencia existente entre las variables (Salvador Figueras, M. 2003)], y a continuación su inercia acumulada. En el resto de la ventana de texto sólo aparecen los datos de los ejes seleccionados (horizontal y vertical): los elementos activos más contributivos a la dimensión (eje) elegida, junto con su posición (*Position*), la contribución del elemento a la inercia de la dimensión, en %, (*Contribution*) y la contribución de la dimensión a la inercia del elemento, en %, (*Quality*) (Bouchet-Valat, M. 2015).

Si se busca determinar un sentido de causalidad en la relación *Contribution* – *Quality* [correlación con el factor o dimensión de un elemento (Navarro Gómez, L. 1983)], hay que comparar la serie de los valores de *Quality* con la de los valores de *Contribution*. Pero como una fuerte Contribution no implica una *Quality* elevada, esto hace que la utilización de las *Qualities* sea delicada y se hace necesario el análisis de *Position*, *Contribution* y *Quality* para poder dar una interpretación correcta y precisa a los factores o dimensiones (Navarro Gómez, L. 1983).

Se observa que las dimensiones o factores obtenidos más representativos del conjunto son el 1, 2 y 3, con una inercia entre el 12 y el 13%, las dos primeras dimensiones sólo recogen el 25% de la inercia acumulada.



Para mayor información sobre el agrupamientojerárquico ascendente del campo *Asbtract* leer el informe completo en el anexo 4.

Los términos más contributivos en la zona negativa del eje 1 son:

	Position	Contribution (%)	Quality (%)
right	-1.6	1.43	5.9
th	-1.7	0.31	5.7
back	-1.4	0.25	6.0
ideas	-1.2	0.23	6.1

Los grupos activos en la zona negativa del eje 1:

	Position	Contribution (%)	Quality (%)
Cluster 11	-1.236	5.5954	6.0603
Cluster 12	-0.025	0.2704	1.5111
Cluster 13	-0.182	0.0734	0.1089
Cluster 15	-0.018	0.0468	0.1037
Cluster 1	-0.295	0.0259	0.0647
Cluster 6	-0.177	0.0158	0.0331
Cluster 5	-0.204	0.0155	0.0404
Cluster 9	-0.112	0.0071	0.0152
Cluster 2	-0.226	0.0061	0.0397
Cluster 4	-0.051	0.0011	0.0023

Los documentos más alejados de la zona negativa del eje 1:

	Position	Quality (%)
scopus_abstract.txt.249	-1.24	6.020
scopus_abstract.txt.247	-1.23	6.033
scopus_abstract.txt.143	-0.38	0.321

scopus_abstract.txt.269	-0.31	0.203
scopus_abstract.txt.306	-0.29	0.065
scopus_abstract.txt.152	-0.23	0.236
scopus_abstract.txt.355	-0.23	0.040
scopus_abstract.txt.32	-0.20	0.040
scopus_abstract.txt.297	-0.18	0.095
scopus_abstract.txt.67	-0.18	0.109
scopus_abstract.txt.98	-0.18	0.109
scopus_abstract.txt.230	-0.18	0.057
scopus_abstract.txt.44	-0.18	0.074
scopus_abstract.txt.112	-0.18	0.033
scopus_abstract.txt.93	-0.16	0.121

Los términos más contributivos de la zona positiva del eje 1 son:

	Position	Contribution (%)	Quality (%)
triggers	13.42	29.520	85.6
transformation	6.86	11.913	83.5
theoretical	8.18	9.966	85.5
evidence	1.48	3.148	81.2
four	4.07	2.964	84.1
computing	4.66	2.267	81.6
studied	3.60	1.742	85.3
systematic	2.29	1.646	83.5
pharmaceutical	1.73	1.641	84.1
included	2.66	1.585	80.6
lung	1.69	1.573	8.4
contrast	3.24	1.561	85.2
fully	2.94	1.414	85.1
literature	1.84	1.318	84.5
areas	1.12	1.187	54.4
main	2.14	1.027	80.6
termed	3.24	0.781	85.2
articles	2.69	0.646	84.9
evolutionary	2.69	0.644	82.3
examined	2.69	0.644	82.3
regulatory	1.31	0.617	80.7
pediatric	1.35	0.601	7.8
pulmonary	1.71	0.478	7.9
regeneration	1.89	0.427	8.0
influences	1.89	0.425	8.0
injury	1.05	0.347	7.8
children	1.17	0.246	5.1
enhance	1.05	0.232	7.8
repair	0.98	0.216	7.7
sleep	1.89	0.213	8.0
diseases	0.30	0.177	5.9
division	1.50	0.167	7.9
childhood	1.25	0.139	7.5
priority	1.24	0.137	7.7
vascular	0.81	0.136	4.9
epigenetic	0.56	0.117	6.8

oriented	0.92	0.101	7.6
aberrant	0.73	0.079	6.9
summarizes	0.65	0.068	2.4
diverse	0.60	0.064	7.1
reviewed	0.60	0.063	7.1
heart	0.57	0.063	3.2
define	0.59	0.062	7.3
held	0.58	0.060	2.7
experience	0.55	0.059	7.0
experts	0.54	0.057	7.1

Los grupos activos de la zona positiva del eje 1 son:

	Position	Contribution (%)	Quality (%)
Cluster 3	6.74	85.3140	86.3231
Cluster 14	1.57	8.1165	8.5334
Cluster 7	0.91	0.4959	0.8995
Cluster 8	0.17	0.0126	0.0360
Cluster 10	0.02	0.0036	0.0068

Los documentos más alejados de la zona positiva del eje 1 son:

	Position	Quality (%)
scopus_abstract.txt.190	7.13	63.721
scopus_abstract.txt.174	6.45	70.525
scopus_abstract.txt.377	1.58	8.326
scopus_abstract.txt.376	1.57	8.514
scopus_abstract.txt.45	0.91	0.900
scopus_abstract.txt.88	0.44	0.737
scopus_abstract.txt.144	0.37	0.541
scopus_abstract.txt.223	0.34	0.234
scopus_abstract.txt.287	0.28	0.098
scopus_abstract.txt.240	0.28	0.220
scopus_abstract.txt.358	0.27	0.234
scopus_abstract.txt.310	0.23	0.070
scopus_abstract.txt.373	0.23	0.125
scopus_abstract.txt.205	0.22	0.062
scopus_abstract.txt.60	0.21	0.224

Los términos más contributivos de la zona negativa del eje 2:

	Position	Contribution (%)	Quality (%)
lung	-5.06	14.66	75.3
pediatric	-4.23	6.08	76.0
pulmonary	-5.29	4.77	76.1
regeneration	-5.85	4.24	76.5
influences	-5.84	4.22	76.5
injury	-3.32	3.58	77.0
children	-3.94	2.88	57.8
enhance	-3.33	2.40	77.3
repair	-3.12	2.26	76.7

diseases	-1.04	2.21	71.4
sleep	-5.84	2.11	76.6
division	-4.67	1.69	76.9
childhood	-3.92	1.43	73.6
vascular	-2.56	1.42	49.4
priority	-3.87	1.39	75.9
epigenetic	-1.83	1.30	72.7
oriented	-2.92	1.06	77.0
summarizes	-2.24	0.85	29.0
aberrant	-2.35	0.85	71.5
held	-1.98	0.73	31.7
diverse	-1.95	0.70	74.3
reviewed	-1.94	0.70	76.2
define	-1.93	0.69	77.8
heart	-1.84	0.68	33.1
experience	-1.80	0.65	74.5
experts	-1.79	0.65	76.9
areas	-0.78	0.60	26.5
included	-0.72	0.12	5.9

Los grupos activos de la zona negativa del eje 2 son:

	Position	Contribution (%)	Quality (%)
Cluster 14	-4.712	75.5775	76.4570
Cluster 4	-0.555	0.1335	0.2670
Cluster 10	-0.078	0.0547	0.0994
Cluster 13	-0.141	0.0455	0.0650
Cluster 15	-0.013	0.0246	0.0525
Cluster 9	-0.135	0.0107	0.0220
Cluster 8	-0.140	0.0085	0.0236
Cluster 7	-0.060	0.0022	0.0038

Los documentos más alejados de la zona negativa del eje 2 son:

	Position	Quality (%)
scopus_abstract.txt.377	-4.76	75.53
scopus_abstract.txt.376	-4.66	75.32
scopus_abstract.txt.117	-0.81	2.54
scopus_abstract.txt.113	-0.77	1.52
scopus_abstract.txt.268	-0.64	0.55
scopus_abstract.txt.358	-0.64	1.30
scopus_abstract.txt.103	-0.55	0.27
scopus_abstract.txt.185	-0.48	0.57
scopus_abstract.txt.175	-0.48	0.81
scopus_abstract.txt.127	-0.45	0.99
scopus_abstract.txt.373	-0.40	0.39
scopus_abstract.txt.25	-0.39	1.09
scopus_abstract.txt.53	-0.37	0.35
scopus_abstract.txt.205	-0.36	0.17
scopus_abstract.txt.295	-0.36	0.34

Los términos más representativos de la zona positiva del eje 2 son:

	Position	Contribution (%)	Quality (%)
triggers	5.18	4.57	13
right	2.19	2.90	11
transformation	2.62	1.80	12
theoretical	3.18	1.56	13
th	2.42	0.64	11
evidence	0.60	0.54	14
back	1.91	0.51	12
ideas	1.73	0.46	12
four	1.58	0.46	13
computing	1.81	0.35	12
systematic	0.93	0.28	14
studied	1.42	0.28	13
pharmaceutical	0.70	0.28	14
contrast	1.29	0.26	13
fully	1.17	0.23	14
literature	0.75	0.23	14
main	0.85	0.17	13
termed	1.29	0.13	13
regulatory	0.56	0.12	15
evolutionary	1.09	0.11	14
examined	1.09	0.11	14
articles	1.08	0.11	14

Los grupos activos en la zona positiva del eje 2 son:

	Position	Contribution (%)	Quality (%)
Cluster 3	2.550	12.68412	12.34907
Cluster 11	1.718	11.23855	11.71222
Cluster 12	0.018	0.13467	0.72419
Cluster 1	0.490	0.07445	0.17922
Cluster 6	0.120	0.00762	0.01535
Cluster 5	0.090	0.00315	0.00793
Cluster 2	0.035	0.00015	0.00095

Los documentos más alejados de la zona positiva del eje 2 son:

	Position	Quality (%)
scopus_abstract.txt.190	2.83	10.053
scopus_abstract.txt.174	2.34	9.268
scopus_abstract.txt.249	1.72	11.649
scopus_abstract.txt.247	1.71	11.646
scopus_abstract.txt.306	0.49	0.179
scopus_abstract.txt.143	0.44	0.428
scopus_abstract.txt.144	0.30	0.357
scopus_abstract.txt.269	0.30	0.195
scopus_abstract.txt.152	0.29	0.370
scopus_abstract.txt.51	0.20	0.096
scopus_abstract.txt.2	0.20	0.139
scopus_abstract.txt.296	0.19	0.100

scopus_abstract.txt.60 0.19 0.183
scopus_abstract.txt.240 0.19 0.098
scopus_abstract.txt.211 0.19 0.091

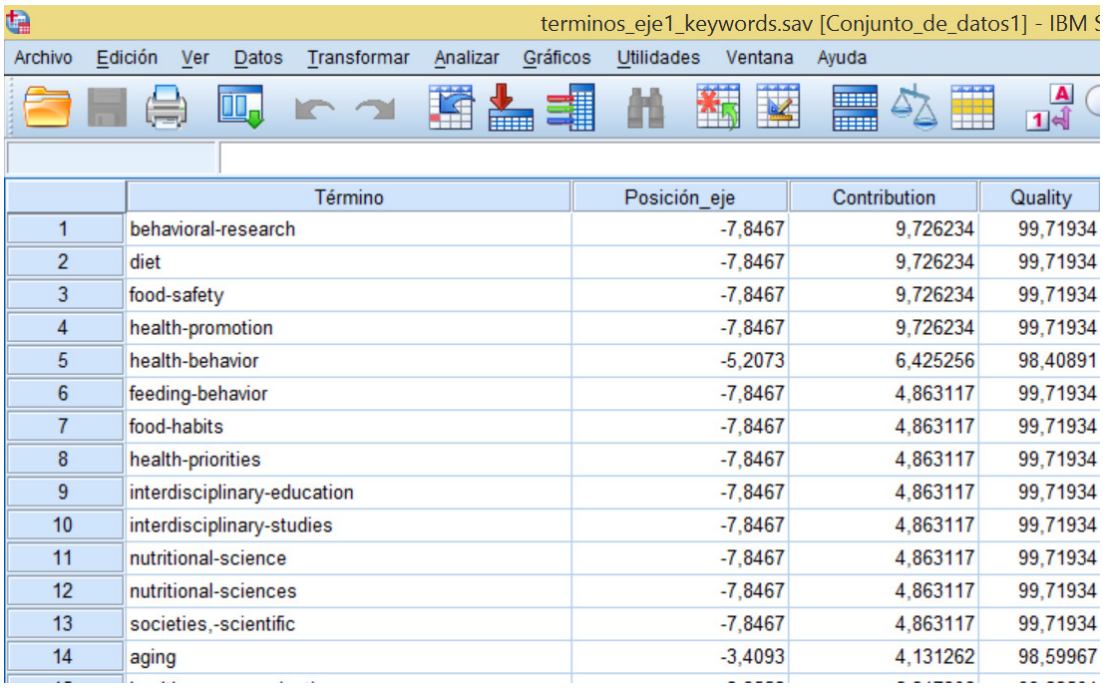
II.- Regresión Lineal Simple de la relación *Contribution* – *Quality* de los términos y *clusters* obtenidos en los análisis de correspondencias. Caso concreto: dimensión 1 del campo *Index Keywords*

Como se ha mencionado previamente, una fuerte *Contribution* no implica una *Quality* elevada, esto hace que la utilización de las *Qualities* sea delicada y se hace necesario el análisis de *Position*, *Contribution* y *Quality* para poder dar una interpretación correcta y precisa a los factores o dimensiones (Navarro Gómez, L. 1983). En esta sección se ha tratado de corroborar esta afirmación comprobando si existe una relación lineal entre las variables *Contribution* y *Quality* para el corpus utilizado en este trabajo

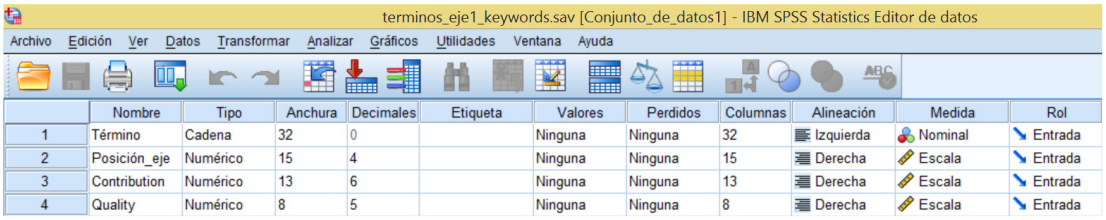
Para ello, se ha aplicado la técnica RLS a las variables *Contribution* y *Quality* obtenidas tras el análisis de correspondencias de los términos y *clusters*. En concreto, la RLS se ha aplicado en el caso del eje 1 (dimensión 1) del campo *Index Keywords*, tanto para los 50 términos más contribuyentes como para los 15 *clusters* obtenidos tras el *Hierarchical clustering*. El resto de casos (eje 2 y ambos ejes del campo *Abstract*) se analizarían de forma similar a la mostrada a continuación.

II.1 Construcción de la tabla de datos

Lo primero es introducir los datos en SPSS. En el caso de los términos tenemos:

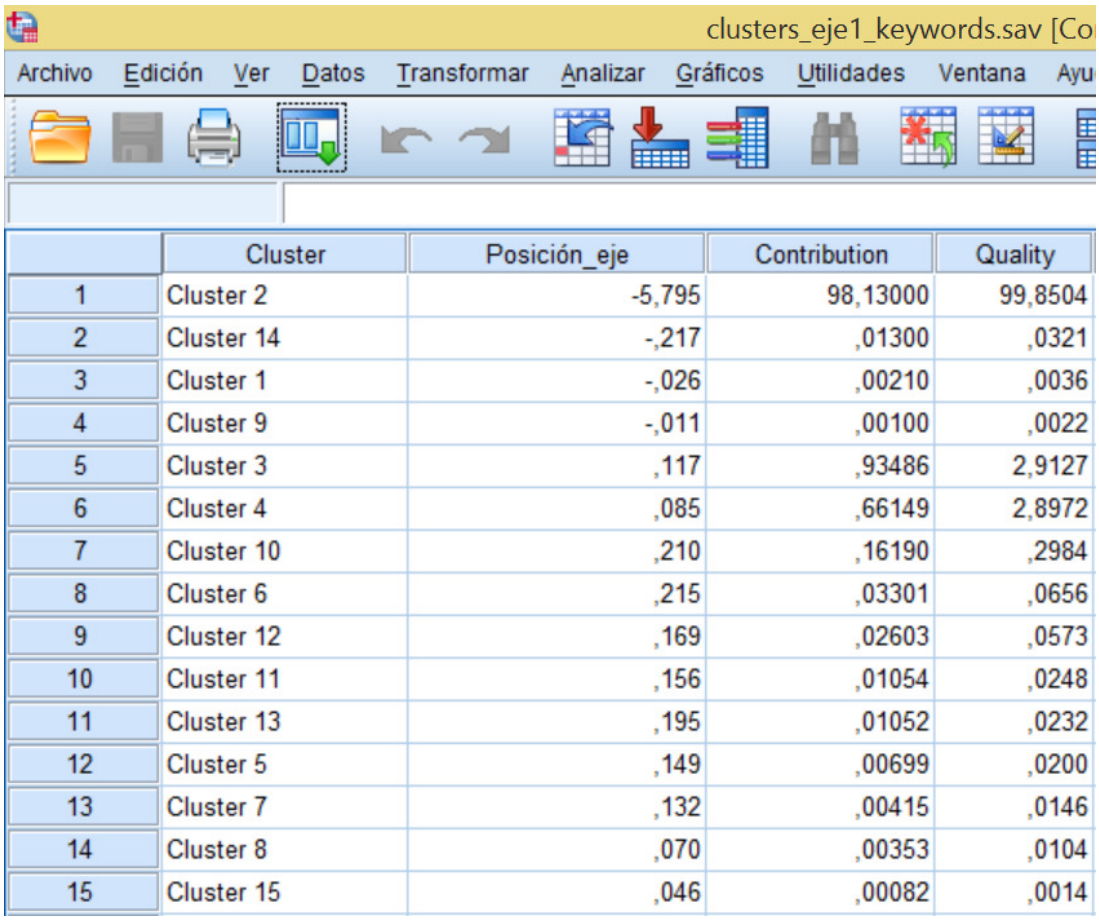


	Término	Posición_eje	Contribution	Quality
1	behavioral-research	-7,8467	9,726234	99,71934
2	diet	-7,8467	9,726234	99,71934
3	food-safety	-7,8467	9,726234	99,71934
4	health-promotion	-7,8467	9,726234	99,71934
5	health-behavior	-5,2073	6,425256	98,40891
6	feeding-behavior	-7,8467	4,863117	99,71934
7	food-habits	-7,8467	4,863117	99,71934
8	health-priorities	-7,8467	4,863117	99,71934
9	interdisciplinary-education	-7,8467	4,863117	99,71934
10	interdisciplinary-studies	-7,8467	4,863117	99,71934
11	nutritional-science	-7,8467	4,863117	99,71934
12	nutritional-sciences	-7,8467	4,863117	99,71934
13	societies,-scientific	-7,8467	4,863117	99,71934
14	aging	-3,4093	4,131262	98,59967



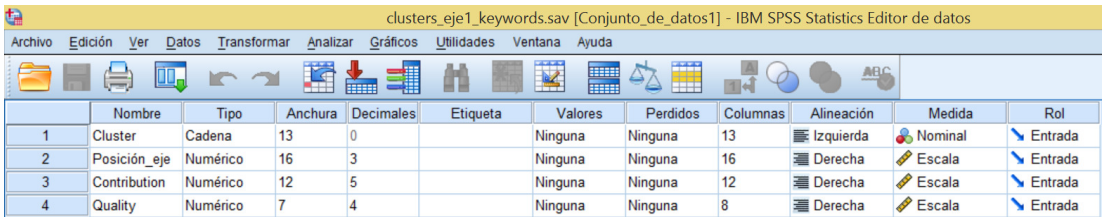
	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Término	Cadena	32	0		Ninguna	Ninguna	32	Izquierda	Nominal	Entrada
2	Posición_eje	Numérico	15	4		Ninguna	Ninguna	15	Derecha	Escala	Entrada
3	Contribution	Numérico	13	6		Ninguna	Ninguna	13	Derecha	Escala	Entrada
4	Quality	Numérico	8	5		Ninguna	Ninguna	8	Derecha	Escala	Entrada

En el caso de los *clusters* tenemos:


clusters_eje1_keywords.sav [Co

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayu

	Cluster	Posición_eje	Contribution	Quality
1	Cluster 2	-5,795	98,13000	99,8504
2	Cluster 14	-,217	,01300	,0321
3	Cluster 1	-,026	,00210	,0036
4	Cluster 9	-,011	,00100	,0022
5	Cluster 3	,117	,93486	2,9127
6	Cluster 4	,085	,66149	2,8972
7	Cluster 10	,210	,16190	,2984
8	Cluster 6	,215	,03301	,0656
9	Cluster 12	,169	,02603	,0573
10	Cluster 11	,156	,01054	,0248
11	Cluster 13	,195	,01052	,0232
12	Cluster 5	,149	,00699	,0200
13	Cluster 7	,132	,00415	,0146
14	Cluster 8	,070	,00353	,0104
15	Cluster 15	,046	,00082	,0014


clusters_eje1_keywords.sav [Conjunto_de_datos1] - IBM SPSS Statistics Editor de datos

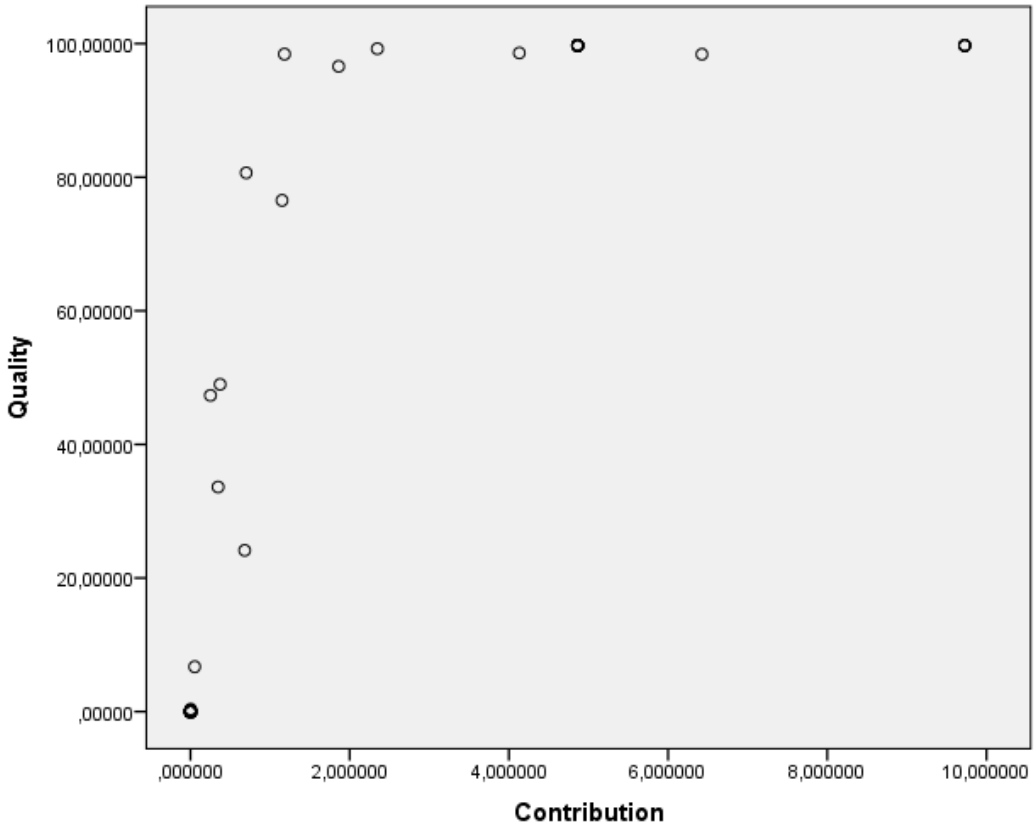
Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Cluster	Cadena	13	0		Ninguna	Ninguna	13	Izquierda	Nominal	Entrada
2	Posición_eje	Numérico	16	3		Ninguna	Ninguna	16	Derecha	Escala	Entrada
3	Contribution	Numérico	12	5		Ninguna	Ninguna	12	Derecha	Escala	Entrada
4	Quality	Numérico	7	4		Ninguna	Ninguna	8	Derecha	Escala	Entrada

En ambos casos la variable independiente es *Contribution* y la variable dependiente es *Quality*.

II.2 Análisis previo de linealidad

Antes de aplicar la RLS, es necesario comprobar si tiene sentido aplicar un modelo lineal para la relación de ambas variables. Para ello, tanto el diagrama de dispersión como el coeficiente de correlación lineal de Pearson son de gran utilidad. En el caso de los términos tenemos:

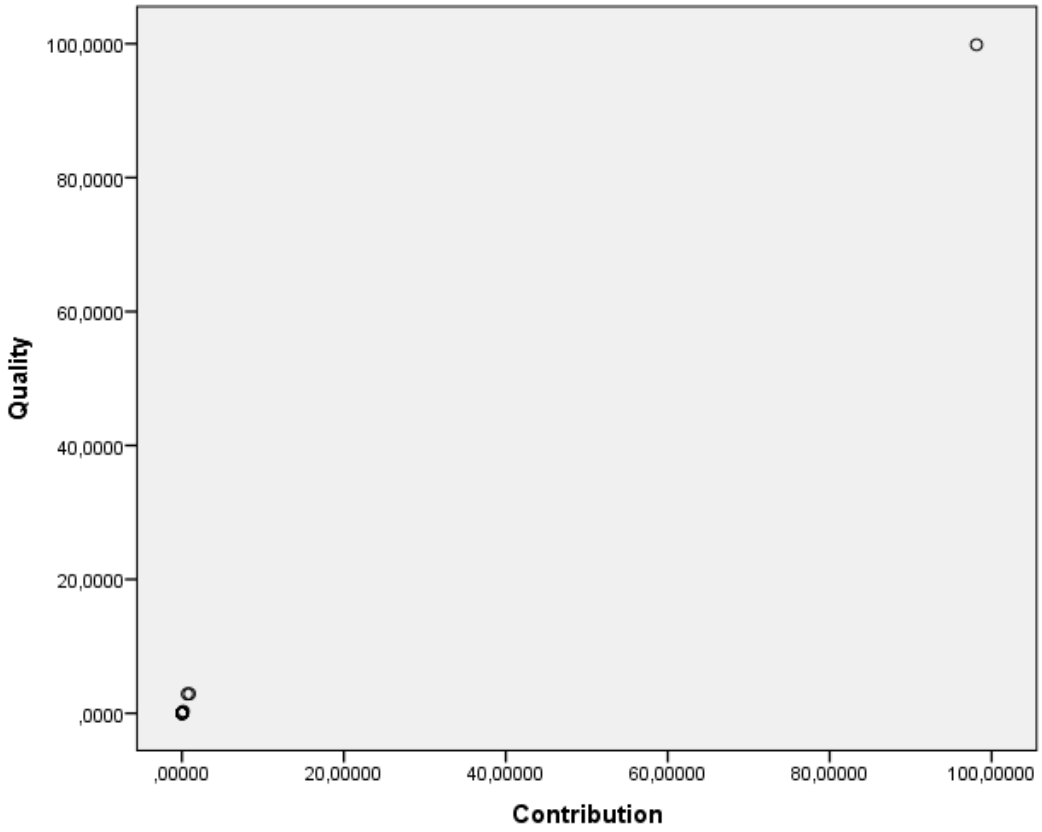


Correlaciones

		Contribution	Quality
Contribution	Correlación de Pearson	1	,816**
	Sig. (bilateral)		,000
	N	50	50
Quality	Correlación de Pearson	,816**	1
	Sig. (bilateral)	,000	
	N	50	50

**. La correlación es significativa al nivel 0,01 (bilateral).

En el caso de los *clusters* tenemos:



Correlaciones

		Contribution	Quality
Contribution	Correlación de Pearson	1	1,000**
	Sig. (bilateral)		,000
	N	15	15
Quality	Correlación de Pearson	1,000**	1
	Sig. (bilateral)	,000	
	N	15	15

**. La correlación es significativa al nivel 0,01 (bilateral).

En el caso de los términos, el coeficiente de Pearson es alto ($\rho = 0.816$) y, por tanto, aboga por una relación lineal entre ambas variables. Si nos fijamos en el gráfico de dispersión, se puede comprobar que la relación se aproxima a una lineal a trozos, es decir, la relación entre las variables se ajustaría mejor a dos rectas definidas en distintos intervalos que a una única recta. Aún así, se ha realizado el ajuste al modelo RLS.

En el caso de los *clusters*, el coeficiente de Pearson es máximo ($\rho = 1.000$) y, por tanto, aboga por una relación lineal entre ambas variables. Si nos fijamos en el gráfico de dispersión, se puede comprobar que la relación se aproxima a una lineal, aunque la mayoría de puntos están concentrados en torno al (0,0) y no parece haber suficiente información en rangos intermedios. Aún así, se ha realizado el ajuste al modelo RLS.

II.3 Aplicación de la RLS

Tras aplicar la RLS con el método *Introducir* y conservando residuos y valores pronosticados para la posterior validación del modelo, los resultados son los siguientes.

En el caso de los términos tenemos:

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Contribution[b]	.	Introducir

a. Variable dependiente: Quality

b. Todas las variables solicitadas introducidas.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,816[a]	,665	,658	27,17076625	,261

a. Variables predictoras: (Constante), Contribution

b. Variable dependiente: Quality

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	70388,712	1	70388,712	95,345	,000[b]
	Residual	35436,026	48	738,251		
	Total	105824,738	49			

a. Variable dependiente: Quality

b. Variables predictoras: (Constante), Contribution

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	15,777	4,581		3,444	,001
	Contribution	12,514	1,282	,816	9,764	,000

a. Variable dependiente: Quality

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típica	N
Valor pronosticado	15,7772856	137,4897919	40,1343208	37,90124452	50
Residual	-37,77044678	67,88104248	0E-8	26,89208436	50
Valor pronosticado tip.	-,643	2,569	,000	1,000	50
Residuo típ.	-1,390	2,498	,000	,990	50

a. Variable dependiente: Quality

En el caso de los *clusters* tenemos:

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Contribution[b]	.	Introducir

a. Variable dependiente: Quality

b. Todas las variables solicitadas introducidas.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	1,000[a]	,999	,999	,7550072	1,131

a. Variables predictoras: (Constante), Contribution

b. Variable dependiente: Quality

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	9227,576	1	9227,576	16187,713	,000[b]
	Residual	7,410	13	,570		
	Total	9234,987	14			

a. Variable dependiente: Quality

b. Variables predictoras: (Constante), Contribution

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	,317	,202		1,569	,141
	Contribution	1,015	,008	1,000	127,231	,000

a. Variable dependiente: Quality

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típica	N
Valor pronosticado	,317858	99,878326	7,080927	25,6731881	15
Residual	-,3164577	1,9090359	0E-7	,7275431	15
Valor pronosticado tip.	-,263	3,615	,000	1,000	15
Residuo típ.	-,419	2,529	,000	,964	15

a. Variable dependiente: Quality

En el caso de los términos, el modelo estimado es *Quality = 15.777 + 12.514 · Contribution*. Es decir, por cada unidad que aumente o disminuya la *Contribution*, la *Quality* aumenta o disminuye proporcionalmente en 12.514 unidades. La bondad del ajuste, dada por el coeficiente de determinación es $R^2 = 0.665$, lo que nos indica que el modelo estimado explica el 66.5 % de la variabilidad que se observa en los datos. Respecto a la hipótesis nula de que no hay relación lineal entre ambas variables, el p-valor de la pendiente (0.000) indica que la hipótesis se rechaza, es decir, que no se puede considerar que no haya relación lineal entre las variables. Por último, la varianza residual (media cuadrática residual) es de 738.251.

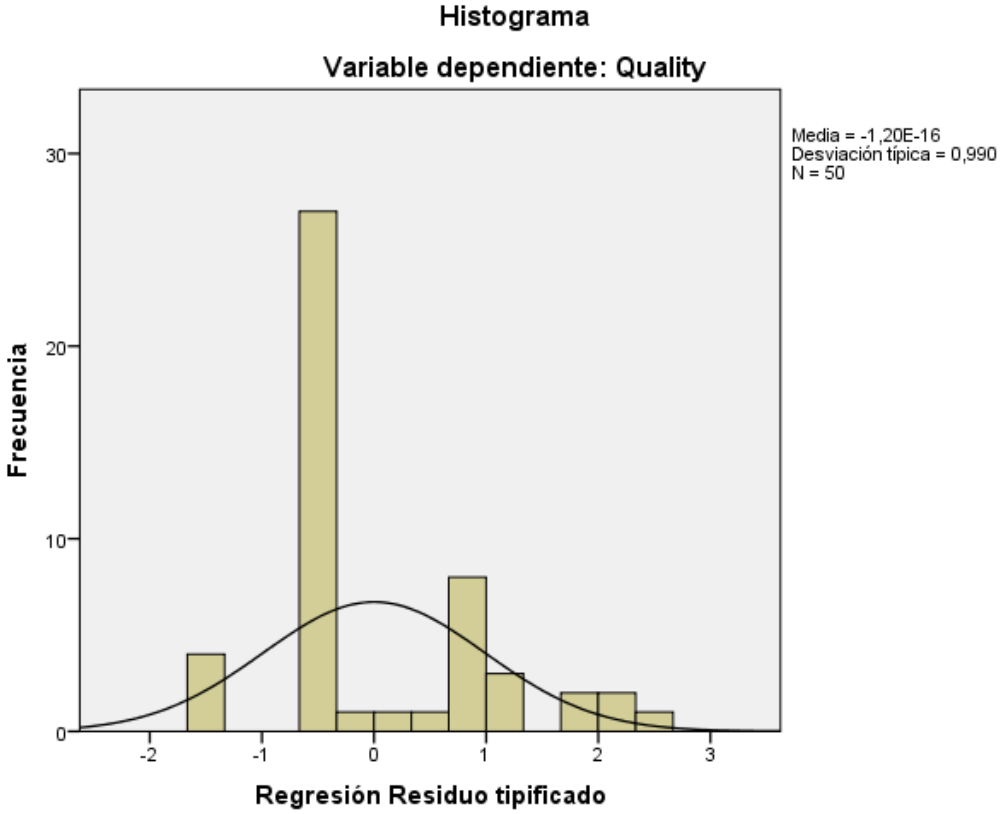
En el caso de los *clusters*, el modelo estimado es *Quality = 0.317 + 1.015 · Contribution*. Es decir, por cada unidad que aumente o disminuya la *Contribution*, la *Quality* aumenta o disminuye proporcionalmente en 1.015 unidades. La bondad del ajuste, dada por el coeficiente de determinación es $R^2 = 0.999$, lo que nos indica que el modelo estimado explica el 99.9 % de la variabilidad que se observa en los datos, es decir, un ajuste muy bueno. Respecto a la hipótesis nula de que no hay relación lineal entre ambas variables, el p-valor de la pendiente (0.000) indica que la hipótesis se rechaza, es decir, que no se puede considerar que no haya relación lineal entre las variables. Por último, la varianza residual (media cuadrática residual) es de 0.570.

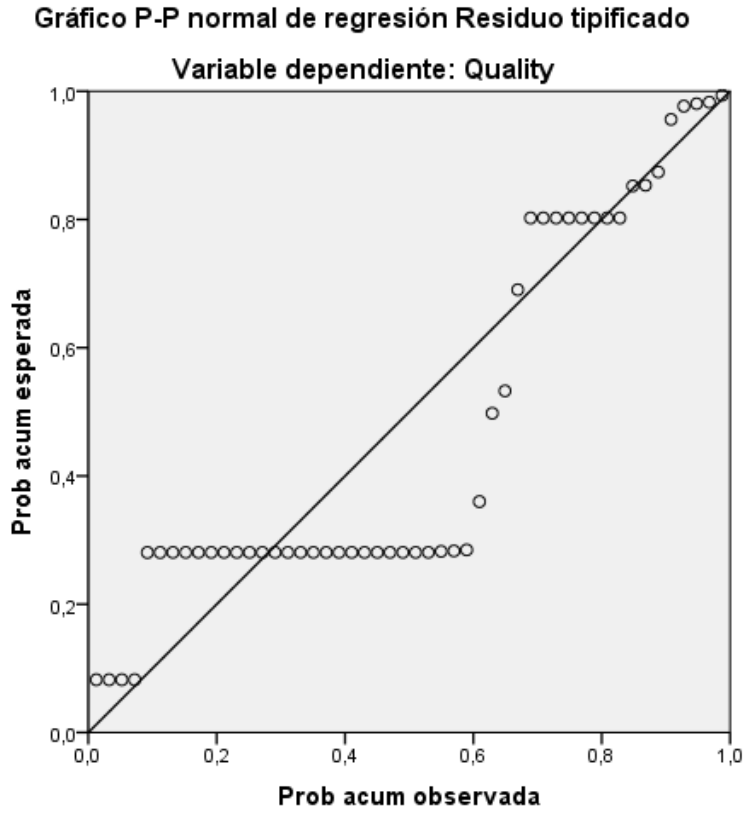
II.4 Validación del modelo

De las cinco hipótesis que se deben cumplir para el correcto funcionamiento de la RLS, la linealidad ha sido comprobada previamente y la homogéneidad se asume cierta porque el modelo tiene término constante tanto en el caso de los términos como en el de los *clusters*. El resto, se han analizado y los resultados se muestran en las siguientes subsecciones.

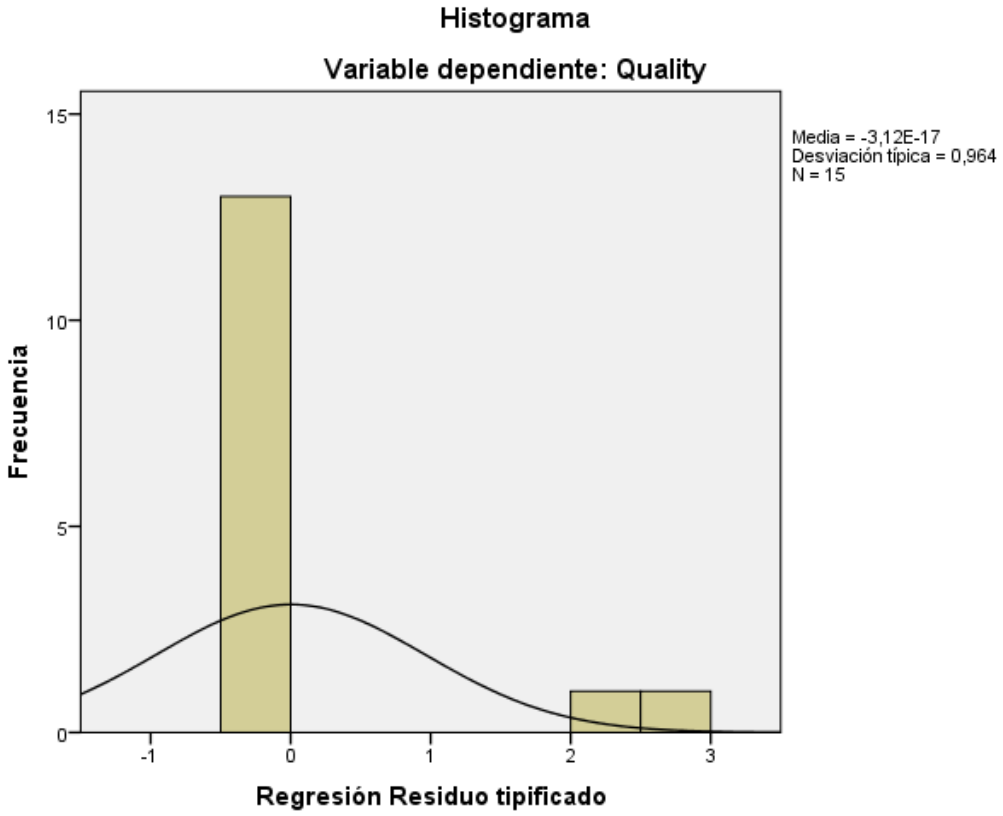
II.4.1.- Normalidad

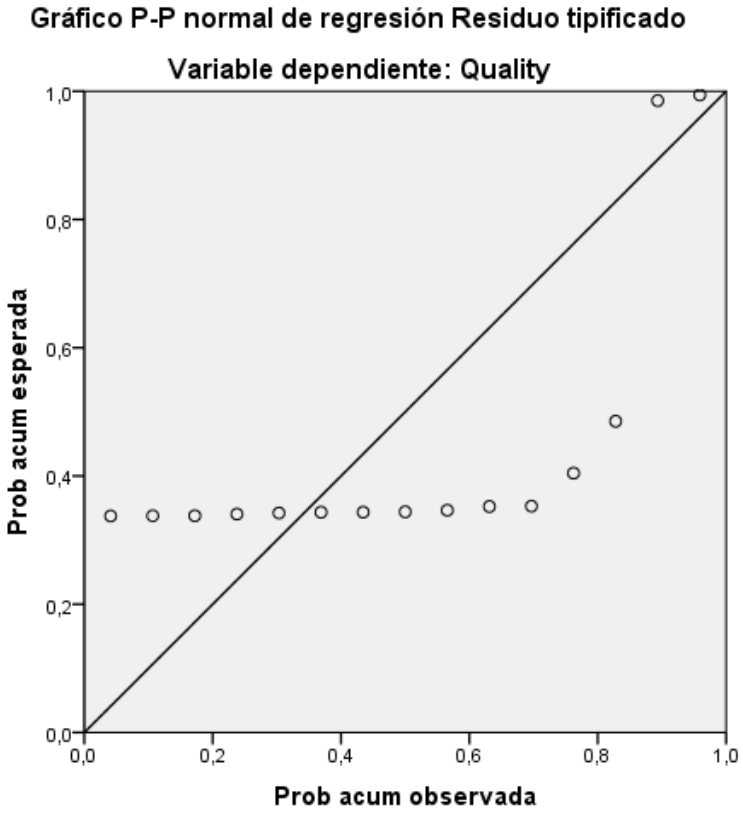
La normalidad se ha analizado mediante el histograma y el gráfico P-P normal de regresión. En el caso de los términos tenemos:





En el caso de los *clusters* tenemos:





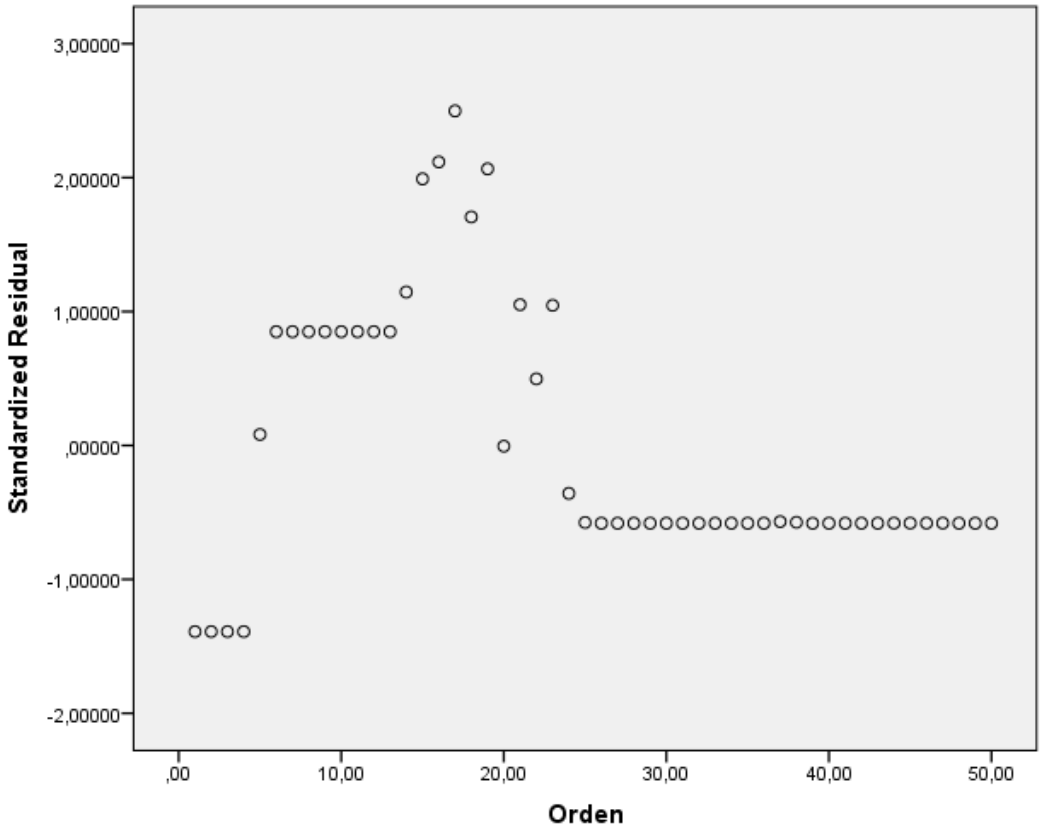
En el caso de los términos, ambos gráficos indican que el grado de normalidad de los datos no es demasiado alto, ya que el histograma se parece poco a la campana de Gauss y la probabilidad acumulada esperada respecto a la probabilidad acumulada observada no se aproxima mucho a la recta Y=X.

En el caso de los *clusters*, ambos gráficos indican que el grado de normalidad de los datos no es demasiado alto, ya que el histograma se parece poco a la campana de Gauss y la probabilidad acumulada esperada respecto a la probabilidad acumulada observada no se aproxima mucho a la recta Y=X.

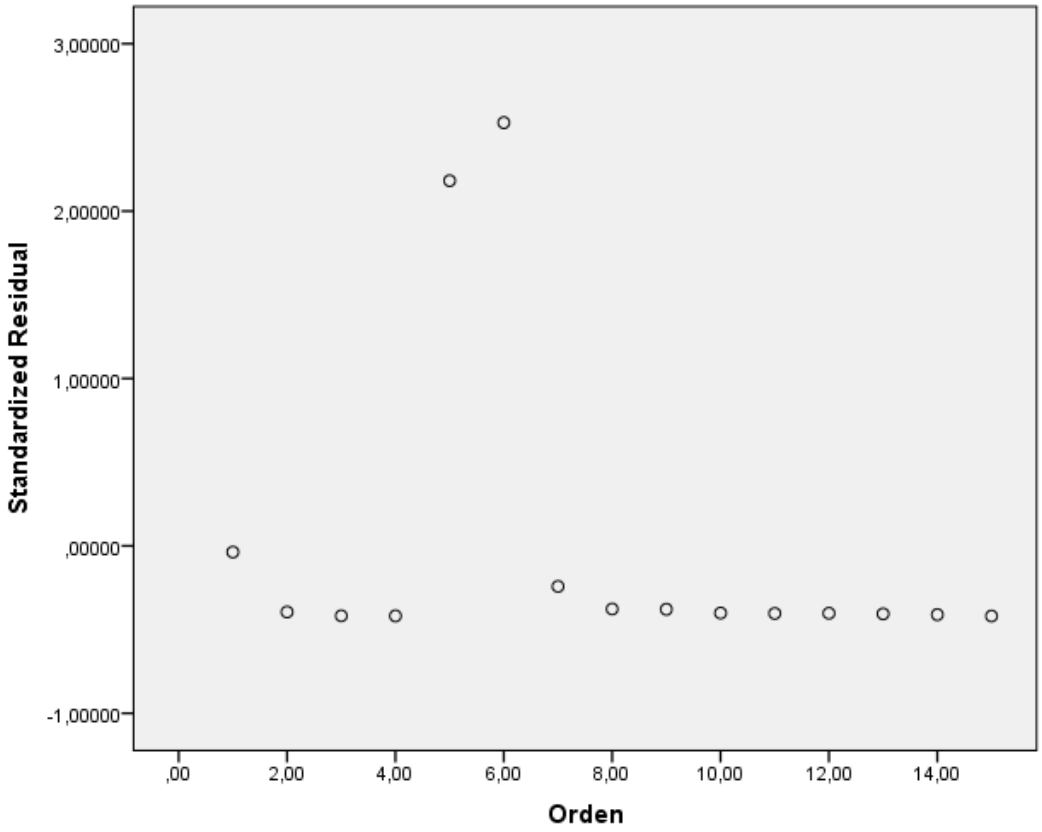
II.4.2.- Independencia

La independencia se ha analizado mediante el estadístico de Durbin-Watson mostrado anteriormente (entre 1 y 2 se puede descartar la dependencia de los datos) y mediante el gráfico de los residuos frente al orden temporal (si no hay patrones de comportamiento se puede descartar la dependencia).

En el caso de los términos, el estadístico de Durbin-Watson es 0.261 y el gráfico es el siguiente:



En el caso de los *clusters*, el estadístico de Durbin-Watson es 1.131 y el gráfico es el siguiente:



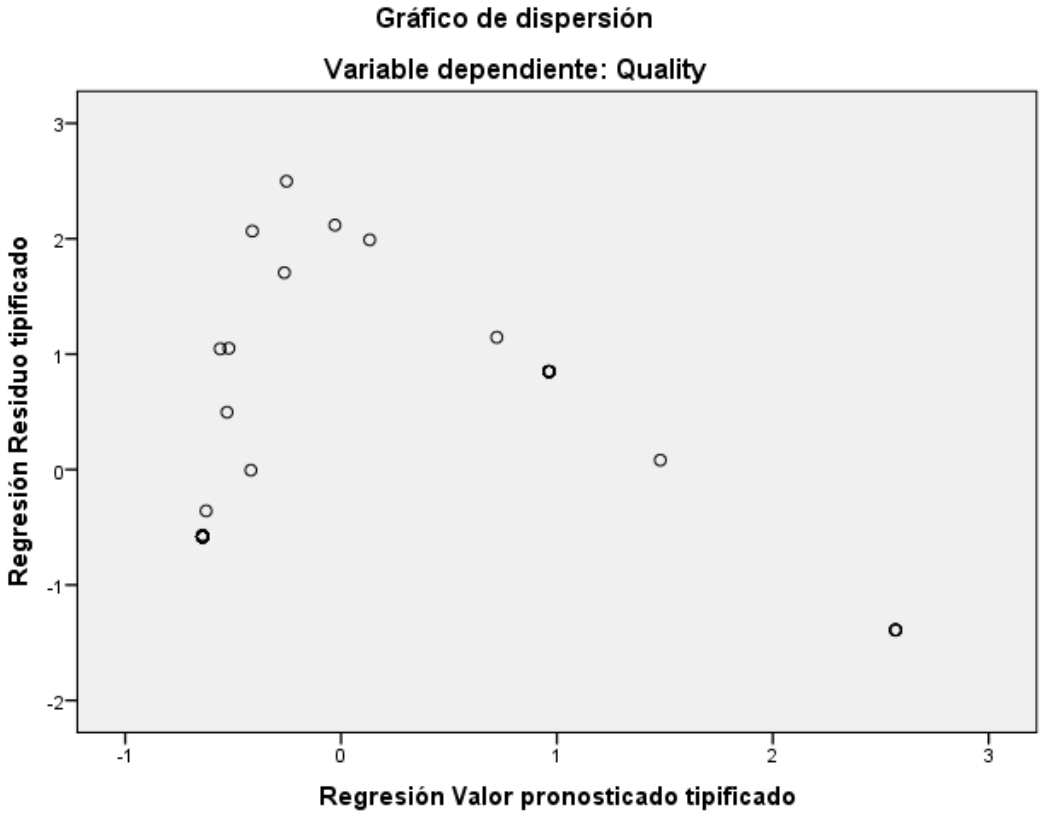
En el caso de los términos, tanto el estadístico (fuera del rango [1,2]) como el gráfico de los residuos frente al orden temporal (se aprecian ciertos patrones de comportamiento lineales), indican que no se puede descartar la dependencia de los datos y, por tanto, no se garantiza la condición de independencia de los datos.

En el caso de los *clusters*, el estadístico (dentro del rango [1,2]) indica que se puede descartar la dependencia de los datos. Por otro lado, en el gráfico de los residuos frente al orden temporal se aprecian ciertos patrones de comportamiento lineales y, por tanto, no se garantiza la condición de independencia de los datos.

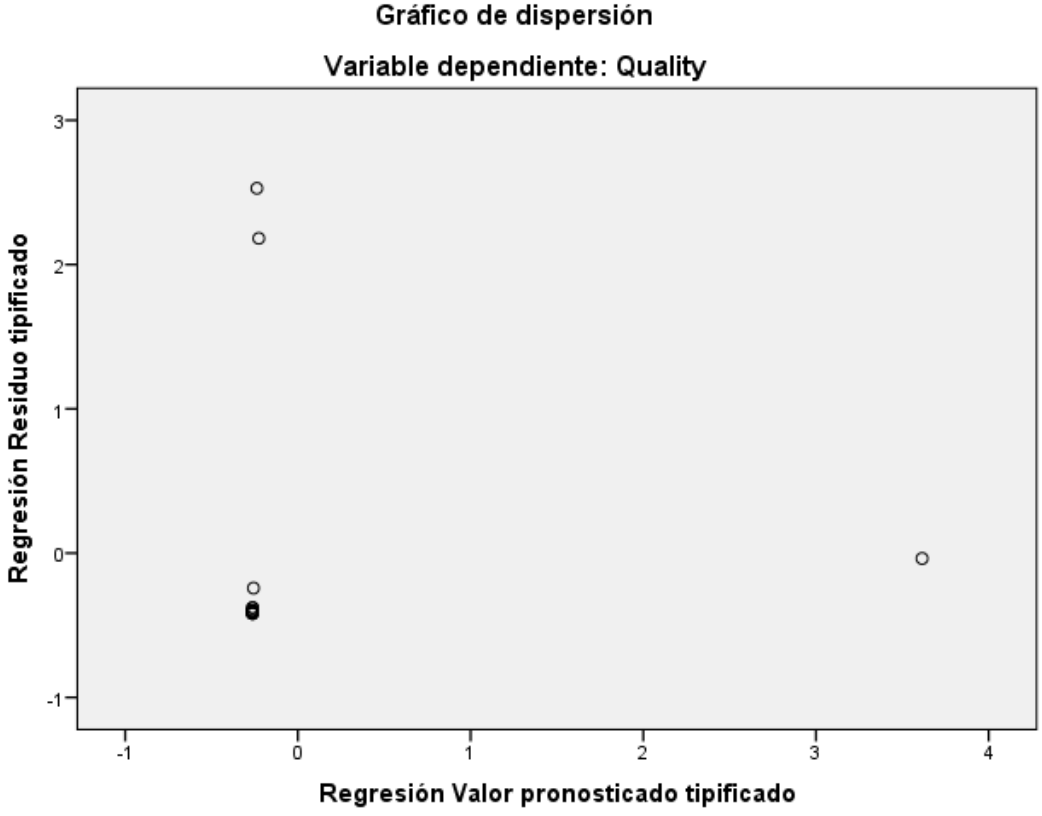
II.4.3.- Homocedasticidad

La homocedasticidad o igualdad de varianzas se ha analizado usando el gráfico de los residuos frente a los pronosticados (si no hay patrones de comportamiento se puede descartar la dependencia).

En el caso de los términos, tenemos:



En el caso de los *clusters*, tenemos:



En el caso de los términos, se aprecian ciertos patrones de comportamiento lineales en el gráfico de los residuos frente a los pronosticados y, por tanto, no se puede garantizar la homocedasticidad de los datos.

En el caso de los *clusters* sucede lo mismo, se aprecian ciertos patrones de comportamiento lineales en el gráfico de los residuos frente a los pronosticados y, por tanto, no se puede garantizar la homocedasticidad de los datos.

II.5 Conclusiones de la RLS

¿Se cumple la afirmación de Navarro Gómez, L. 1983 citada anteriormente o es suficiente con conocer la *Contribution* para poder predecir la *Quality*?

En el caso de los términos, parece claro que no hay una relación lineal única entre ambas variables. Desde el primer análisis de linealidad mediante el gráfico de dispersión, se ha observado que hay una relación lineal a trozos (dependiendo del intervalo en el que se encuentre la variable), pero no se puede encontrar una única recta cuya función modele de forma adecuada la relación entre ambas varables. Esto ha sido corroborado con los resultados obtenidos en la validación del modelo (no se cumplen linealidad, normalidad, independencia y homocedasticidad). Por tanto, en este caso, no es suficiente conocer la variable *Contribution* para conocer los valores que toma la variable *Quality*.

En el caso de los *clusters*, parece que hay una relación lineal entre ambas variables, aunque no se tienen suficientes datos para afirmar esto, sobre todo en rangos intermedios del intervalo. Desde el primer análisis de linealidad mediante el gráfico de dispersión, se ha observado que parece haber una relación lineal, de hecho la bondad del ajuste ha resultado muy alta. En cambio, hay ciertas deficiencias que han sido puestas en evidencia en la validación del modelo. No se cumplen normalidad y homocedasticidad, probablemente debido a la poca cantidad de datos y al vacío de datos en zonas intermedias. La independencia de datos tampoco ha sido garantizada. Por tanto, hay indicios de una relación lineal entre las variables *Contribution* y *Quality* (en zonas extremas), pero no podemos garantizar que exista puesto que no hay suficientes datos.

Por último, puntualizar que se podía haber hecho alguna predicción con los modelos estimados pero no era el objetivo de la RLS en este contexto.

Conclusiones

La estadística lexical y el análisis de datos textuales se encuentran en la base del desarrollo de los denominados sistemas de conocimiento, sistemas expertos fruto de la ingeniería del conocimiento, muy empleados en *Text Mining* .

En el presente trabajo se han aplicado técnicas de minería de textos a un corpus documental de una disciplina científica reciente, *Traslational Research and Personalized Medicine*. Se ha comprobado que el grado de diferenciación del nuevo campo científico es escaso (la variabilidad interclusters es mínima, es decir, escasa variabilidad en temas de investigación). La técnica del análisis factorial de correspondencias empleada ha permitido comprobar esta escasa diferenciación temática al estructurar el conjunto de términos en función de los *clusters,* obtenidos con anterioridad en la clasificación jerárquica ascendente. Como se observa en el diagrama de dos dimensiones, la dispersión de los *clusters* es escasa, tan solo los *clusters*

2 y 15, se alejan de un grupo denso formado por el resto. Se comprueba igualmente que la inercia total de los ejes, factores o dimensiones es baja en todos ellos, es decir, la contribución de cada una de las dimensiones es tan baja que la interpretación de cada una de las dos dimensiones es compleja.

Respecto a la Regresión Lineal Simple aplicada a la relación *Contribution* – *Quality* de términos y *clusters* para un caso concreto, se ha comprobado que para los términos, la relación parece ser lineal a trozos y, para los clusters, hay indicios de haber una relación lineal aunque faltan datos para poder afirmar esto, sobre todo en zonas intermedias. Parece que la relación "cuanto más *Contribution* (peso del elemento, es decir, término o *cluster*, en la dimensión o eje) más *Quality* (peso de la dimensión en el elemento)", se hace más fuerte conforme se reduce el número de casos mediante el agrupamiento en *clusters*. Esto tiene sentido puesto que al reducir la dimensión de los datos nos aproximamos más al número de ejes o dimensiones al que se han reducido los elementos tras el análisis de correspondencias (2 ejes).

Bibliografía

Berzal, F. Clustering jerárquico [cited 1 January 2016]. Available from world wide web: <http://elvex.ugr.es/idbis/dm/slides/42%20Clustering%20-%20Hierarchical.pdf>.

Bouchet-Valat, M. , 2015. Package "RcmdrPlugin.temis" [cited 4 January 2016]. Available from world wide web: <https://cran.r-project.org/web/packages/RcmdrPlugin.temis/RcmdrPlugin.temis.pdf>.

Bouchet-Valat, M., and Bastin, G., 2013. RcmdrPlugin.temis, a Graphical Integrated Text Mining Solution in R. *The R Journal* 5, 188–196.

Cabo Salvador, J. Investigación traslacional. Definición. Objetivos. *gestion-sanitaria.com* [cited 1 January 2016]. Available from world wide web: <http://www.gestion-sanitaria.com/investigacion-traslacional-definicion-objetivos.html>.

Casanova, J.F. Análisis factorial de correspondencias [cited 1 January 2016]. Available from world wide web: <https://www.uam.es/personal_pdi/medicina/casanova/Analisis_Correspondencias.pdf>.

Gamero Estévez, E. Medicina personalizada [cited 1 January 2016]. Available from world wide web: <https://www.upo.es/moleqla/export/sites/moleqla/documentos/Articulo_Destacado_1.pdf>.

Garnier, B. Utilisation d'un outil de statistiques textuelles R.TeMiS (Plugin de R Commander). [cited 1 January 2016]. Available from world wide web: <https://s3-eu-west-1.amazonaws.com/r.temis/Pas+%C3%A0+Pas+sous+R.TeMiS+0.7.2.pdf>.

Greenacre, M., Nenadic, O., and Friendly, M. , 2015. Package "ca" [cited 4 January 2016]. Available from world wide web: <https://cran.r-project.org/web/packages/ca/ca.pdf>.

Guardiola Jiménez, P. Análisis de correspondencias, [cited 1 January 2016]. Available from world wide web: <http://www.um.es/docencia/pguardio/documentos/Tec_Homals.pdf>.

Gutiérrez, R., González, A., Torres, F., and Gallardo, J. A., 1994. Métodos jerárquicos de análisis cluster. In *Técnicas de análisis de datos multivariable. Tratamiento computacional* [cited 1 January 2016]. Available from world wide web: <http://www.ugr.es/~gallardo/pdf/cluster-3.pdf>.

Navarro Gómez, L. Aspectos teóricos y una aplicación práctica del análisis factorial de correspondencias. *Estadística española* 1983, 33–59.

Salvador Figueras, M., 2003. Análisis de Correspondencias [cited 1 April 2016]. Available from world wide web: <http://www.5campus.com/leccion/correspondencias>.

Terrádez Gurrea, M. Análisis de conglomerados [cited 1 January 2016]. Available from world wide web: <http://www.uoc.edu/in3/emath/docs/Cluster.pdf>.

Universidad de Granada, 2016. Máster oficial en investigación traslacional y medicina personalizada (TransMed) [cited 1 January 2016]. Available from world wide web: <http://masteres.ugr.es/transmed/>.

Wehling, M. , 2008. Translational medicine: science or wishful thinking. *Journal of Translational Medicine*, 31.